Social Impact Retrieval: Measuring Author Influence on Information Retrieval

James Lanagan

A dissertation submitted in fulfilment of the requirements for the award of

Doctor of Philosophy (Ph.D.)

to the

Dublin City University
School of Computing

Supervisor: Prof. Alan F. Smeaton

June 2009

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:

ID No:

Date:

ACKNOWLEDGEMENTS

Firstly I would like to thank my supervisor Alan Smeaton for the help he has given me in finding a direction for my research. I must thank him also for the effort he has put in to the balancing act of knowing when to let me follow my own ideas, whilst still providing that direction.
A collection of people helped me through this thesis, all of whom provided me with intelligent insight and the odd push to keep me focussed and confident that the research could be done:

- During the creation of both my initial systems, Hyowon Lee was of immeasurable help in advising and tweaking the ideas that I had for how the systems should look.
- Throughout this thesis Paul Ferguson and Pete Wilkins have both provided valuable knowledge and help in getting through my experiments, providing me with expert judgements, and lastly with perhaps some questionable sporting wisdom.
- Colum Foley, Adam Bermingham, Cathal Gurrin, and Gareth Jones all listened to my thoughts towards the later months of my thesis and helped to push me over the finish line.
- A word of thanks to all those both past and present within the CDVP who have helped me enjoy my PhD experience and in particular: Georgina, Sinéad, Kieran, Mike, Kevin, Phil, Fabrice, Sandra, Neil, Edel, Daragh and Niamh.
- Mike Christel first got me into the information retrieval game, and put me in contact with Alan and the CDVP. For this I am always grateful.
- Mick Cooney for all the advice and direction he has provided thanks to his incredible knowledge of all things statistical or otherwise.

I must also thank most sincerely my parents Geoff and Jane for all they have done for me through my many adventures. Your unwavering support, along with that of my sister Sarah has meant so much. I am truly blessed.
Lastly, I thank my girlfriend Linda for all that she has put up with through these last few years. It has been made easier each day by your constant patience and encouragement and I could not have done this without you. Thank you from the bottom of my heart.

ABSTRACT

The increased presence of technologies collectively referred to as Web 2.0 mean the entire process of new media production and dissemination has moved away from an authorcentric approach. Casual web users and browsers are increasingly able to play a more active role in the information creation process. This means that the traditional ways in which information sources may be validated and scored must adapt accordingly.

In this thesis we propose a new way in which to look at a user's contributions to the network in which they are present, using these interactions to provide a measure of authority and centrality to the user. This measure is then used to attribute an query-independent interest score to each of the contributions the author makes, enabling us to provide other users with relevant information which has been of greatest interest to a community of like-minded users. This is done through the development of two algorithms; AuthorRank and MessageRank.

We present two real-world user experiments which focussed around multimedia annotation and browsing systems that we built; these systems were novel in themselves, bringing together video and text browsing, as well as free-text annotation. Using these systems as examples of real-world applications for our approaches, we then look at a larger-scale experiment based on the author and citation networks of a ten year period of the ACM SIGIR conference on information retrieval between 1997-2007. We use the citation context of SIGIR publications as a proxy for annotations, constructing large social networks between authors. Against these networks we show the effectiveness of incorporating user generated content, or annotations, to improve information retrieval.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The democratisation of information production and publishing processes on the internet today means that the number of sources from which information may have come from is increasing hugely. It is no longer the case that information on a single webpage has come from a single source; many web pages, because of deliberate syndication of information as we get with news, and from direct end-user cut-and-paste replication, contain information from several sources. In addition to this, internet users may now add annotations of many different forms and types. This facility whereby users add information such as annotations is part of the "social web" and varies from users' sharing of bookmarks[1], tagging of multimedia[2] or online interaction, video uploads etc. It is now possible for internet users to add in-context annotations and information to any webpage, and without any form of filtering or authentication. This move from author-centric to community-centric production and publication means that there can no longer be a reliance on the source of information as an indication of the quality, trustworthiness, or value of that information. Accordingly, new metrics for measuring these aspects of a source of information should be devised. In addition, these metrics must take into account not only the source of new information, but the context in which this information is gathered. These metrics should take into account not just the current interactions, but also the interactions with users in the past. It is in this broad area of information management that we focus on in this thesis, in developing and testing ways in which information context becomes as important as information content. Firstly, however, we will introduce a popular topic among internet users, social information retrieval.

1.1 Social Information Retrieval

Creating annotations on existing web content is a form of interacting with other web users and leads to what is called social information retrieval. This broadly describes a

[1] http://delicious.com/
[2] http://www.flickr.com/

new way in which we are able to use the world wide web and its content and is in fact a far more natural way of communicating. By our nature human are a gregarious species who in general will prefer the company of others. The ability to share information and see the contributions of a community, be they people we know or otherwise, redefines the web from a static space filled with single-user entities effectively unaware of each other, to a dynamic space in which users are able to communicate and share opinions about anything, anywhere, any time. Users are no longer constrained to forums or newsgroups, nor by knowledge of specific technologies. Instead, for better or for worse, anyone is now able to find a voice for their arguments and insights.

Casual web users and browsers are increasingly able to play a more active role in the information creation process. This means that the traditional ways in which information sources may be validated and scored must adapt accordingly. Collaborative filtering provided one of the first methods of utilising the interactions of users with a system in order to improve its performance. Tapestry (Goldberg et al., 1992) introduced the idea of using individuals' interactions with an email client to aid in the filtering of email for every user of the email client. By allowing single users to annotate their email, the system incorporates this feedback into its own behaviour. In doing so, Tapestry utilised one of the first instances of community voting. The mechanism which Tapestry relied upon was the explicit annotation and rating of e-mails by the users of the system, as well as a specific method of interaction in order to take advantage of these annotations. This coupled with its use amongst a small and task-orientated group (workers within the same office) made its form of collaborative filtering inapplicable to larger, more web-based communities.

Recommendation systems such as GroupLens (Konstan et al., 1997; Shardanand and Maes, 1995) extended the collaborative filtering idea to larger scale communities of users without the prerequisite of real-world acquaintance. Systems such as these use the ratings provided by the user community to rate the items within its collection. These ratings are then utilised in providing recommendations to users (e.g. say what films to watch); without the system itself having to know anything about the actual content of the items. This type of recommendation is possible without the requirement of user profiles, since the information being used is simply the rating. Allowing the creation of profiles however adds significant advantages such as personalisation, weighting of recommendations based on who has provided these recommendations, and user-based as opposed to item-based recommendation (Balabanović and Shoham, 1997).

Since these systems were developed, large scale creation and use of user-generated content and information has become the norm. Many different internet services are now offered which allow internet users to tag, annotate, create links between and even combine or 'mash-up' exciting websites with no interaction from or reference to the original creator of the underlying documents. Through the use of technologies collectively

referred to as *Web 2.0* the entire process of new media production and dissemination democratised.

1.1.1 User Generated Content and the Social Web

The nature of information sharing has begun to change also. Instead of sharing information on a one-to-one basis such as e-mail, there is an increasing trend towards the one-to-many style of publication. Services such as Twitter[3] and Facebook[4] have seen a huge surge of membership numbers in recent times (Nielsen, 2009).These sites offer users the ability to share information with close friends, family, or the web at large. This information may be in the form of video, audio, images or text, and the amount of content being shared continues to grow.

The content being shared is not necessarily just that which is created by the sharer. Other types of content include the provision of additonal meta-data for online resources such as descriptions or tags. Folksonomy is a portmanteau of the words 'folk' and 'taxonomy', and refers to a taxonomy of terms created by a collection of users. The addition of terms is not regulated by any central authority but instead may be added to any resource (most commonly web-pages) and by any user. It has been noted that organised ontologies may arise from the seemingly chaotic assignment of tags to resources by an uncontrolled and unrestricted user community (Mika, 2007). The combination of folksonomies and content-based image management has become a strong area of research within the Semantic Web (Berners-Lee et al., 2001). Tags added by users may be used to reduce the semantic gap between what a picture represents, and what a simple image processing can perceive (Wu et al., 2006; Mika, 2007). With the addition of tag information, the gap between meaning and representation is reduced to understanding what each and every tag means. This in itself can be a problem due to the lack of controls or consistency check in place when adding tags.

Beyond tags, free-text annotation systems allow for a more descriptive and elaborate form of annotation. The World Wide Web Consortium (W3C), the group responsible for overseeing and directing the implementation of technical standards across the web, has continued developing its own annotation platform, Annotea[5], which is based on existing W3 standards. This development by the leading group in world wide web development is an important justification for research referenced and continued in this thesis. As mentioned, human are gregarious by nature and past research has shown that this is also true of our behaviour on the web. Users will visit web-sites that others have visited in the past, not just by chance but by preference (Freyne et al., 2007). It is around this

[3] http://www.twitter.com
[4] http://www.facebook.com
[5] http://www.w3.org/2001/Annotea/

phenomenon of social interactions among humans over the world wide web, that we have formulated an hypothesis. After surveying the research literature, we have identified an area where there is a gap in the available research. We now describe this hypothesis.

1.2 Thesis Hypothesis

We propose that the interactions and free-text annotations created within the context of the world wide web as a whole may be used to improve the relevance judgements of documents returned in answer to a user's information need. By looking not just at query-dependant measures such as the content of a document, but also at the annotations created on a document, we shall show that the overall 'social impact' of a document may help in fulfilling this need. Other query-independent measures such as profile information of the creators of annotations, as well as the network of annotations themselves help provide more of an insight in to how interesting the community as a whole may find particular documents. Using these features, we aim to provide users with documents which are not only relevant to their information need, but that also use the "wisdom of crowds" to place importance on those documents/items which have been favoured by the community of users as a whole.

Our approach differs from that of either collaborative filtering or content-based retrieval since there is no specific burden placed on the users to provide a rating for any of the documents which they annotate. Rather it uses the natural process of conversation and interaction as a guide to finding those items which have proved the most interesting to the user community. We do store a profile of each of the users of our system, but this profile consists of the interaction and associations that the user creates within their own social network. The work presented in this thesis does not attempt to utilise this profile to find like-minded users, but instead to gauge the importance of a user's contributions to the social network of users as a whole. The approach we present is important as it allows for retrieval which is more social in nature, mimicking the concept of "word-of-mouth" more closely. As we shall see in Chapter 2, people will naturally trust information which comes from a source they know rather than from a strange one. Our approach aims to utilise the idioms of "word-of-mouth" and "voting with your feet"; in a community, if the views of a particular person elicit no response from any other person within the community then these opinions should be considered of little value to the community at large. This leads us to the following hypothesis:

> *"The ranking of documents returned in answer to a user's information need may be improved by incorporating information from the social network of documents' authors, as well as the network of annotations on the documents themselves."*

We believe that with the ever-growing number of contributors and contributions to the world wide web, both in the form of self-contained documents/web-sites, and as meta-data or content on these websites, there is a need to provide measures by which these contributions may be rated and valued. This is the focus of the work presented here.

1.3 Research Objectives

In order to validate our hypothesis, we shall create and examine the characteristics of 3 corpora of varying size and origin. Firstly, we shall discuss two real-world systems built to study the usefulness and potential of user-generated content to aid in information retrieval and in the browsing process. These systems were deployed and focussed on high-profile sporting events which took place within the last 3 years. The novelty of the systems lay in their ability to bring together and combine 3 currently separate aspects of sports recording; viewing, analysis, and discourse. The creation of our first two data-sets was a direct result of this discourse. After examination of these systems, we outline the creation and analysis of our third corpus, built from the citation and author networks of the ACM SIGIR conference proceedings. This community is shown to accurately approximate the community of users which we would expect to find in a 'social web' or internet community. In examining these corpora we aim to satisfy the hypothesis stated above. To do so, we have identified a number of questions which need to be answered in order to provide evidence for the ideas which we have put forward:

1. If users are given the opportunity to annotate documents, will they do so?
    i) Do users find the annotations of others within the community interesting?
    ii) Do users enjoy the additional interaction and social element which is introduced through the use of annotation?
    iii) Do users value the contribution of others?
2. Are the annotations that users create on a 'social web' corpus of use to the user community as a whole?
    i) Can these annotations be leveraged to improve the overall performance of the system in satisfying users' information needs?
    ii) Can we identify specific elements of a user's profile of interactions which are of use in the ordering and ranking of documents to benefit the user?
    iii) Can the processes of "word-of-mouth" and "voting with your feet" be automated?

We shall compare the algorithms which we have developed to current state-of-the-art approaches for authority measurement, showing that they perform as well in providing value to the users of our community. We shall then extend this to consider the contributions of our users to the pool of community knowledge, providing a measure of interest and value for each annotation which is created. In doing so we aim to show that not only are the annotations of others of interest to the community, but they may also be leveraged to improve the browsing, searching and general utility of a corpus to its users.

1.4 Thesis Organisation

This thesis is organised as follows:

Chapter 1: In chapter 1 we provide a brief introduction to our research problem and outline the contents of this thesis.

Chapter 2: In chapter 2 we provide a background of the areas of research and activity which have influenced the direction of this thesis. We look first at the general field of information retrieval, providing a grounding in the techniques used to retrieve and rank documents from a collection. We then look at the field of social network analysis and discuss work on attributing importance to agents within a network, most importantly by using the dynamics of interaction and web of trust which is built between these agents. Lastly we look at the areas of data quality and annotation, helping us to learn the motivation and value behind user-generated content.

Chapter 3: In chapter 3 we consider many of the new methods of interaction amongst internet users collectively called "Web 2.0". We give an overview of the state-of-the-art in Web 2.0 research and applications. Next we outline 2 novel social media systems, SportsAnno and Annoby, which were developed as part of this thesis to help understand the ways in which people create and share information. An analysis of usage and design of these systems is presented, along with the lessons learned from their implementation.

Chapter 4: In chapter 4 we introduce a second corpus of pseudo-annotations based on the citation network of SIGIR proceedings from 1997-2007. We discuss the area of citation analysis and justify our usage of citations as a proxy for user-generated annotations. After doing so, we discuss the collection of this corpus of our extended SIGIR corpus. We present the results of analysis on the characteristics of the author and citation networks of the corpus, drawing parallels between it and the SportsAnno and Annoby corpora

presented in chapter 3. Lastly we present the two algorithms, AuthorRank, A_R, and MessageRank, M_R, which we have created to exploit these networks.

Chapter 5: In chapter 5 we discuss the collection and creation of a ground-truth against which to compare the techniques we have developed in this thesis, as well as other state-of-the-art metrics based on both implicit user-feedback and citation analysis. Statistical analysis is performed on this ground-truth created through user experiments to ensure a level of consistency and agreement. Once this has been completed, we compare the rankings provided by our experts to those of the well-known Google Scholar search engine, and other methods widely used in current research practices.

Chapter 6: In chapter 6 we detail the experiments which have been undertaken to explore the effectiveness and usefulness of the algorithms detailed in the previous chapters. Firstly we describe the systems we have built which allow us to compare the individual features within our algorithms. We then combine these features to take advantage of each of their distinct characteristics. We shall also look at the effectiveness of current state-of-the-art citation analysis algorithms in the SIGIR and Web 2.0 context. Finally we show that the techniques we have developed and trained on our extended SIGIR corpus are indeed of benefit in improving the rankings of documents returned as the result of a query to an information retrieval system.

Chapter 7: Finally, in chapter 7 we summarise our results, suggest extensions to our approach, and describe future work.

CHAPTER II

BACKGROUND & RELATED WORK

In this chapter we provide an overview of the areas of research which have come together to influence and direct the work in this thesis. Each of the areas presented here has had some impact on the hypothesis underlying this thesis. Firstly, we discuss information retrieval and the means by which information may be organised and searched so as to help users find information which is of greatest relevance to their current information need.

Secondly, we discuss network analysis and the creation of networks of users. Through this, we are able to study and understand more clearly group dynamics of the users whom we shall rank and classify individually later in the thesis. Social network analysis is the specific aspect of this field that is of greatest relevance to our own work.

Authority and trust provide a means by which to assign some measure of importance to members of a random user community that is able to write and annotate objects freely. A trust metric is one which is able to rank users not just by what they write but also by their standing within the community or network of users as a whole. Authority and trust thus play a role in determining content ranking later in the thesis.

There is now renewed interest in measuring and using the quality of information created through tasks like annotation; the basis for which comes from data quality itself. We discuss the formal theory of data quality as well as approaches to measuring it which provide a theoretical grounding to the algorithms we present later in this thesis.

Finally, annotation itself provides a means for users to interact with media in general.

In our case, annotation allows users to interact explicitly with media within the systems we have created. Annotation also forms a basis for the pseudo-annotations created within our SIGIR corpus that is described later. We discuss the role of annotation in the physical and digital domains, as well as the many diverse uses and interpretations of annotations as a whole.

2.1 Information Retrieval

Information Retrieval (IR) has its roots in the 1950s when an increasing number of scientists and researchers began to realise that the speed at which information could be indexed and catalogued was falling behind the speed at which new information was being created. Automatic methods of indexing and retrieving information had become necessary. Luhn (1958) pioneered the concept of using the terms within a text document to index it, allowing the frequency of the term to dictate its importance within the document, as well as its relevance when searching. Luhn (1957) states that *"It is hereby proposed that the frequency of word occurrence in an article furnished a useful measurement of word significance"*. This work formed the basis of many of the 'best match' retrieval strategies discussed Section 2.1.2.

Concrete work on the use and limitations of automatic information retrieval began with the Cranfield Experiment (Cleverdon, 1967) which formalised a methodology for experimentation. The SMART system (Salton, 1971b) providing the first working IR system to test these experimental methodologies. Having provided a methodology and system, work advanced throughout the 1960s and 1970s on developing new ways of accessing and indexing information. The two most significant advances made were the Vector Space Model (Salton et al., 1975) and the Probabilistic Model (Robertson and Sparck Jones, 1976), building on the work of Luhn and helping to alleviate some of the weaknesses of the Boolean model which had been developed earlier.

The advent of the World Wide Web meant that the creation and dissemination of information began to grow exponentially. This growth meant that methods for finding relevant information were a necessity. Up until then most information access had been confined to collections of written material on separate and local networks. With the vast amounts of available information on which to test and train new algorithms/approaches, there was a lack of comparable results against which decisions about retrieval and indexing success could be made. In 1992, research regained a more directed and structured framework with the inception of the the Text REtrieval Conference[1] (TREC) series. Established by the National Institute of Standards and Technology (NIST) (a US governmental organisation), the aim of TREC has been and still is to promote research in

[1] http://trec.nist.gov/

the field of information retrieval whilst providing a corpus of documents and an evaluation infrastructure on which to perform this research. Along with the document corpus, NIST also provided topics and metrics to perform evaluation calculations. TREC is now in its 19th year (2009) and participation continues to increase.

2.1.1 An Information System

In our modern society we are now comfortable with the expectation that information is available from any source and consequently may be in any form. As stated in Section 2.4.1.1, information is built from data, which in turn is made up of signals. In considering an information system, we shall restrict ourselves to describing a text-retrieval system. While the over-arching steps described below are not specific to text-retrieval, many of the pre-processing steps are.

The move from paper to digital media over recent years has allowed for the restructuring of content search into ways which were not possible beforehand. The structure of the documents being searched is no longer a restriction as the digital medium means that information can be stored in many different forms, and therefore accessed in many different ways. The main purpose of any information retrieval system however, is to aid a user in satisfying their information need. This is achieved by finding relevant sources from within a document collection. Before information may be retrieved, a number of steps must be performed. These steps are illustrated in Figure 1.

Before any information may be indexed, it may need to be retrieved from sources outside the system. This act of *document gathering* or *corpus creation* can be performed in many ways. The usual way in web-based systems is through a *crawl* of web pages following the hyper-links between the pages and downloading documents which are to be included into the document corpus. This crawl is performed by a *spider*, aptly named since it is the World Wide Web (WWW) which originally gave rise to the hyper-linking and therefore crawling phenomenon. Though the World Wide Web and subsets of the Web are commonly used as corpora, any type of information may be used. For other forms of information such as books or images, different acquisition methods need to be employed; in the case of manuscripts or other hand-written materials, Optical Character Recognition (OCR) may be used after digitisation.

2.1.1.1 Pre-Processing and Indexing

In order to improve the retrieval performance and efficiency of a system, a number of pre-processing stages must be performed on the document corpus. The basic unit of retrieval and indexing within a text-based Information Retrieval system is the 'term' (word). While this allows for a more fine-grained retrieval process, it also means that the

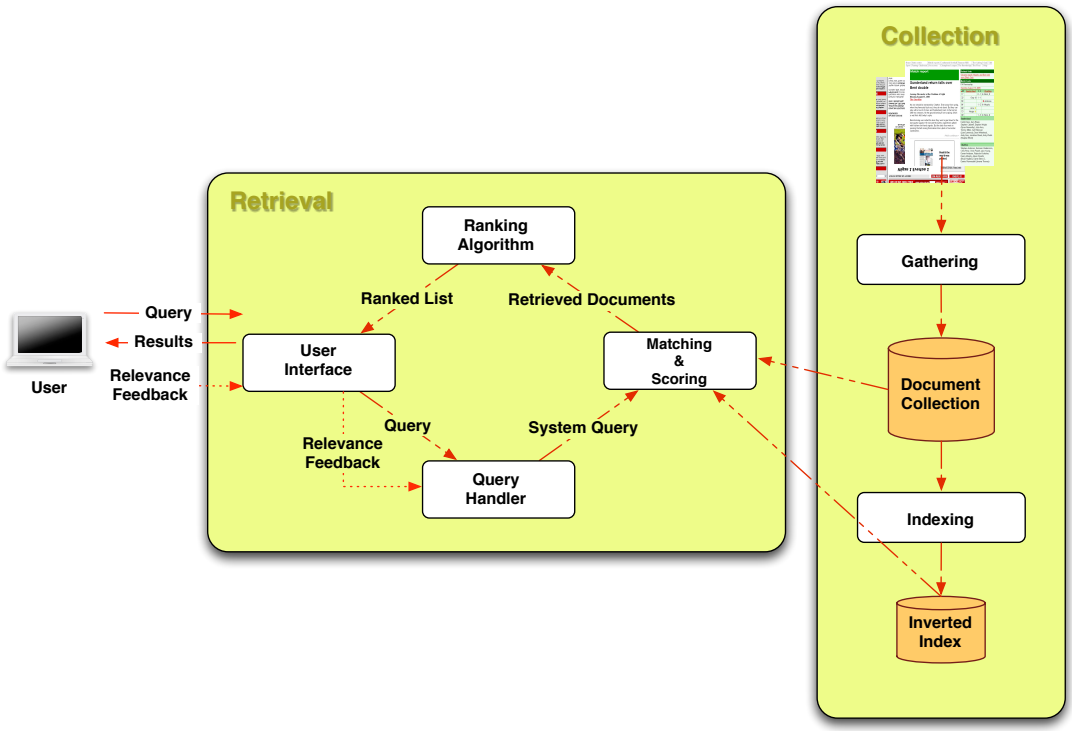


Figure 1: A typical information system illustrating the necessary steps for storage and retrieval of information.

storage of information as raw documents becomes highly inefficient. Another problem is that of redundancy between terms, as well as errors and anomalies created through conflation (synonyms, transliteration, mis-spelling etc.). In order to counter this, two major techniques have been developed to minimise the storage and indexing time required for a document corpus, stopping and stemming.

Stopping is the process of removing words which are of low discriminative value, occurring in the vast majority of documents within the corpus. The obvious purpose of a unit of retrieval is to discriminate between relevant and irrelevant documents, and so those which do not are of little use. Luhn (1957) found that terms which occur very often, as well as those which occur very rarely (perhaps due to spelling mistakes), are of little discriminative value. He called this the 'resolving power' of the word. The discriminative value or resolving power of words within a corpus is in fact an example of Zipf's Law (Zipf, 1949) which states:

$$frequency(f) \times rank(r) = constant \tag{1}$$

Zipf's law has been found to hold for all manner of distributions from various areas of life, distributions as seemingly dissimilar as city populations and alphabetic letter occurrences (Zipf, 1949).

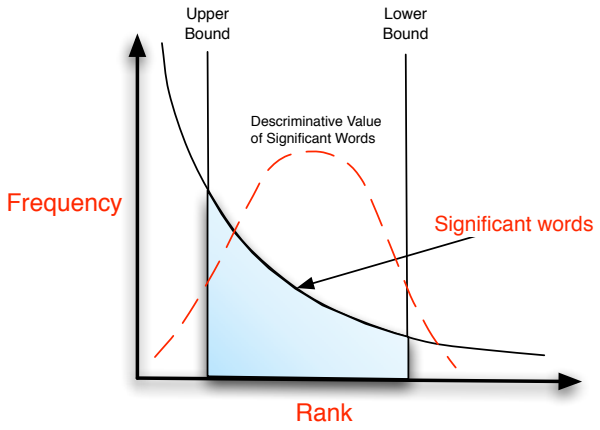


Figure 2: The discriminative value of terms demonstrates Zipf's Law. (Van Rijsbergen, 1979)

By examining the discriminative (or resolving) power of terms which occur within a corpus, Baeza-Yates et al. (1999) found that the size of a corpus may be reduced by as much as 40%. This reduction has implications for storage requirements as well as a positive impact on query response times. Within an English language corpus, common stop-words would be 'a', 'the' and 'am'. These words may be compiled into a stop-word list, examples of which are easily found on the web[2]. It may also be necessary to augment these lists with domain-specific stop-words; 'patient' and 'suffers' would be commonly occurring terms within a medical corpus, for example[3].

Stemming is the process where by words are reduced to their entomological root, reducing the corpus through the removal of plurals, conjugations, etc. As an example, let us take the words "bake", "baking" and "bakery". Through the use of a stemming algorithm such as the commonly used Porter stemming algorithm (Porter, 1980), these three words are reduced to their common root, "bak". This not only has the advantage of reducing the index size, but also means that relevant documents may be found for more queries as those queries can use any of the forms of the word. As with stopping, an additional benefit of reduced query response times may also be observed.

Lastly, the documents are transformed into a more machine-readable format. This is done by transforming each document into a "bag-of-words" representation. In doing so the structure of the original document is lost and we make the assumption that the semantic meaning of the document may still be recovered from the terms within the document.

Once these pre-processing steps have been completed, the collection must be indexed.

[2]http://snowball.tartarus.org/

[3]It should be noted that many commercial search engines no longer perform stopping, as this reduces the effectiveness of the index in returning exact-match phrasal queries (e.g. a search for the band "The The"). Stopping may be performed on the query itself when no phrases are present.

Witten et al. (1999a) note that while there are many ways to index a collection *"in applications involving text, the single most suitable structure is an inverted index"*. The index is "inverted" as it uses terms within documents as a key to locating documents, as opposed to using the documents themselves as the key into the collection. Each document within the collection is first given a unique ID. Each item within the index (referred to as a *'posting'*) then gives the term which is indexed, as well as the ID of each of the documents in which the term occurs. Techniques such as *d-gaps* or *run lengths* (as used in video and audio encoding) may be used to further compress the size of the index (Witten et al., 1999b).

2.1.1.2 Searching and Issuing Queries

Research has shown that while there is a vast amount of information available to web users, the way in which searches are performed is rather limited. Users can have only a vague idea of the information they require when beginning to search and they use a broad approach to iteratively improve their query. The average query made to a text search engine is just over 2 words in length (Jansen et al., 2000; Silverstein et al., 1999). Indeed Silverstein et al. (1999) notes that only 12.6% of queries contain more than 3 words. With such a small query, ambiguity amongst terms as well as the vagueness of the information need can lead to vast numbers of documents being returned. Figure 3 shows the initial search page of the well known Google search engine[4].

Once a query is submitted to the search engine it is handled using the steps shown in Figure 1. Firstly stopping and stemming are performed on the query so it resembles the terms within the inverted-index. The query is then issued to the index and a list of documents which are believed to be relevant is returned. This is done in accordance with the *retrieval model* being used. Before returning this list to the user, the documents are ranked in order of relevance. (In the case of Google, this ranking was originally based on the PageRank algorithm discussed in Section 2.1.3.1.) This ranked list of documents is then returned to the user. The ranking itself is important as Silverstein et al. (1999) noticed that *"surprisingly, for 85% of the queries only the first result screen is viewed"*.

2.1.2 Retrieval Strategies

The way in which this query is handled, and the techniques used to find relevant documents within the corpus varies. In this section we will discuss the three classical retrieval models used to satisfy users' information needs; the Boolean Model, the Vector Space Model, and the Probabilistic Model. All these methods use term distributions within the documents to decide on the relevance to a query. There are however several other

[4] http://www.google.com

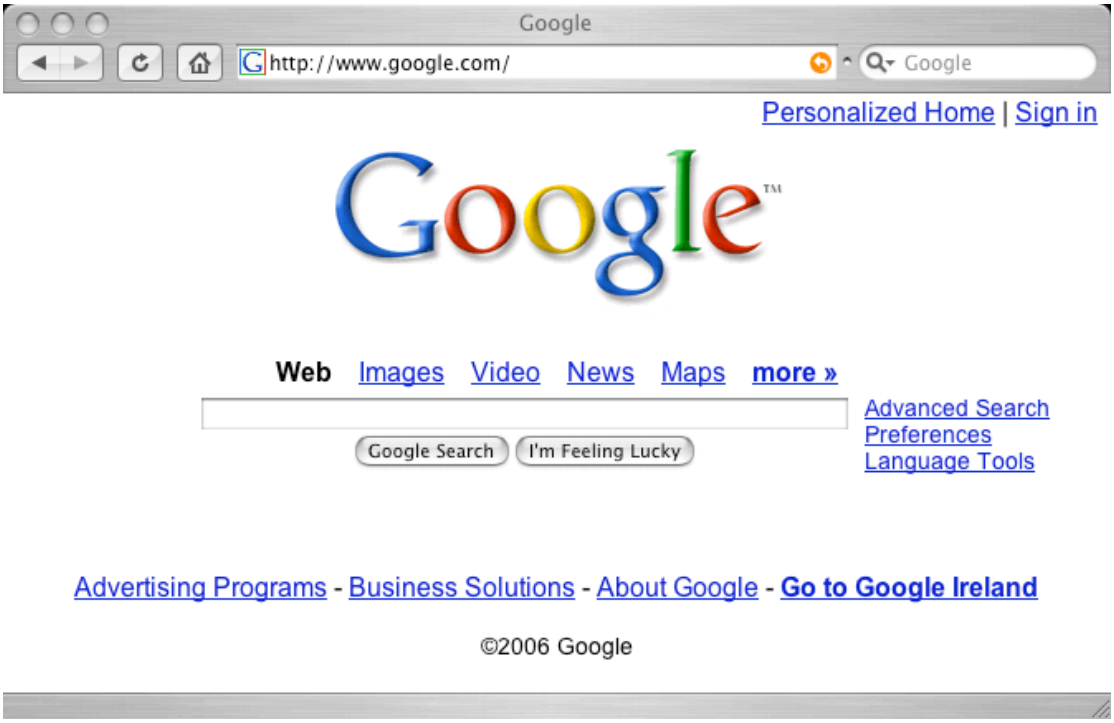


Figure 3: The Google search interface.

ways in which to measure the relevance of a document, such as the link structure of the document and its neighbours. These techniques are discussed in Section 2.1.3.

2.1.2.1 The Boolean Model

For a long time the bulk information retrieval research was focussed on the Boolean model of retrieval. This was because most practical retrieval was performed by trained intermediaries such as librarians. These intermediaries were able to convert the needs of users into a form which was machine-understandable, could choose an appropriate information repository, and were able to extract abstracts or summaries from the raw data sources.

Boolean retrieval is based on Boolean logic and consists of the operators AND, OR, and NOT. Any combination of these may be used to create increasingly complex queries. This is generally done in an interactive fashion, with the user refining the query by increasing its complexity. The approach is referred to as *set-theoretic* because it deals with the sets of documents which contain the query terms, and the intersection, union, and complements of those sets.

An example of a Boolean query may be seen in Figure 4. Here the information need is for documents which are relevant to the three terms; 'Elbow', 'Band' and 'Mercury'. It is important to note that in the example query given, the documents which will be

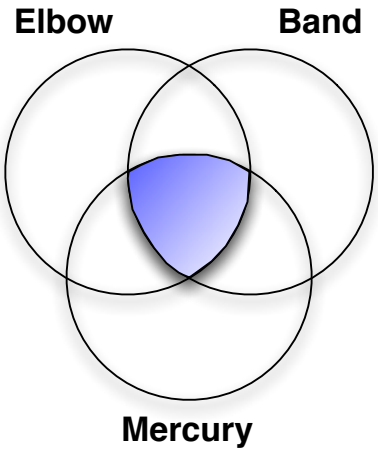


Figure 4: Venn diagram of a Boolean query.

returned (those corresponding to the shaded area in Figure 4) are those which contain all the desired query terms.

Boolean retrieval does not allow for near or partial matches, and there is no way of weighting the terms in the query. There is also no way of ranking the results returned in order of relevance; documents are either relevant or non-relevant to the query. A consequence of this is that there is no real control over the number of documents which are returned for a query. These weaknesses have meant that Boolean search is better suited to more experienced users (Cleverdon, 1988).

2.1.2.2 Vector Space Model

The Vector-Space model, proposed by Salton et al. (1975), represents queries and documents as vectors, with an orthogonal dimension for each term in the collection. Since not every term is in every document, this can lead to sparse vectors. By using the model we are able to compare the similarity of a query, Q against any document, D_i, in the collection. In order to do so we must define a weighting scheme for the terms within the document collection, and also a similarity function with which to compare query and document.

There are many ways in which the similarity of the query Q and a document D_i may be measured. One way is to compare the inner-product of the two respective vectors, but the most commonly used measure is the cosine of the angle between the query and document vectors, defined as:

$$sim(D_i, Q) = \frac{D_i \cdot Q}{|D_i| \times |Q|} \tag{2}$$

or:

$$sim(D_i, Q) = \frac{\sum_{t \in Q} w_{t,D_i} \times w_{t,Q}}{\sqrt{\sum_{t \in Q} w_{t,D_i}^2 \times \sum_{t \in Q} w_{t,Q}^2}} \tag{3}$$

where $w_{t,d}$ is the weight assigned to term t in document D_i and $w_{t,Q}$ is the weight of term t in query Q. This has the nice property that $sim(D_i, Q)$ will be 1 if the document and query are identical, and 0 if they are orthogonal.

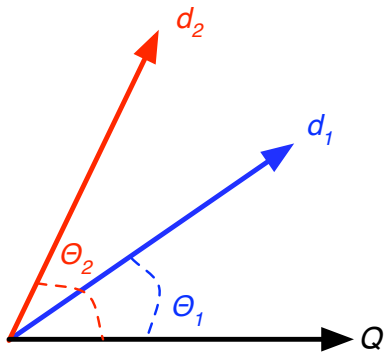


Figure 5: Cosine difference between query and corpus document vectors

When weighting terms within a collection, the easiest weight to apply is a simple binary weight $[0, 1]$, denoting the presence or absence of a term in a document. Statistics on the frequency of occurrence of a term within a document, its *term frequency* tf, may be used to add weight to frequently occurring terms. As stated in Section 2.1.1.1, terms may also be used to discriminate between relevant and non-relevant documents. Terms that occur in fewer documents are often more valuable than those which occur frequently throughout the collection. The *inverse document frequency* (idf) (Sparck Jones, 1972), or collection frequency is a commonly used measure of the prominence and distribution of terms in a collection.

$$idf_t = log\left(\frac{N}{n_t}\right) \tag{4}$$

where N is the total number of documents in the collection, and n_t is the number of documents in the collection that contain the term t.

In a comparison of different weighting schemes, Salton and Yang (1973) found that a combination of these two values worked well:

$$w_{t,d} = tf_{t,d} \times idf_t \tag{5}$$

This basic $tf_{t,d} \times idf_t$ weighting has the drawback that it can give additional weight to terms which occur frequently within longer documents, making longer documents more relevant. Since it is commonly accepted that relevance should be independent of document length, it is therefore necessary to perform some type of normalisation so as to remove the influence of document length. Research by Salton and Buckley (1988) led to the following normalisation which incorporates the maximum within-document term frequency, $maxtf$:

$$w_{ij} = \frac{tf_{ij}}{maxtf_{ij}} \times log\left(\frac{N}{df_j}\right) \tag{6}$$

Singhal et al. (1996) extended this work further, incorporating *pivoting* to compensate for the favouring of long documents in retrieval.

The main benefit of the vector-space model, as with all best-match methods when compared to the Boolean model, is that it does not require an exact match to query terms for a document to be returned. This leads to levels of similarity in the returned documents, a fact which may be exploited to yield ranked lists of results. Unlike the Boolean model the number of results can also be limited to, say just the first 10, 100, etc. These advantages, as well as the ease of implementation of the algorithms required, have led to the vector-space model becoming very popular and highly used in modern information retrieval systems.

2.1.2.3 Probabilistic Model

The probabilistic approach to retrieval attempts to return documents which are of probable relevance to a user's information need. Unlike the vector-space model which returns documents based on the similarity of the document to a query, this model returns documents based on the probability that they will be relevant to the query. The model was first proposed by Maron and Kuhns (1960) to help solve the so-called "library problem". The model aims to predict whether a document, D, is relevant, R, to a query, Q, with probability $P(R|Q, D)$. Robertson and Sparck Jones (1976) developed the underlying research in an attempt to provide some theoretical grounding to the process of retrieval. While the vector-space model takes into account frequencies of occurrence, its underlying mathematics are quite ad-hoc. For example, the scores assigned to documents are not probabilities, but rather estimated measures of relevance. Subsequent to the development of the probabilistic model for retrieval, the *Probabilistic Ranking Principle* (Robertson, 1977) was proposed which states:

> If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to

> the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data.

The most basic form of the model is the *Binary Independence Model* (Robertson and Sparck Jones, 1976) which makes the assumption that term occurrence is stochastically independent, and that a document is either relevant or non-relevant to a query. The probability of relevance is computed based on certain attributes or features of a document, typically the terms or phrases within the document. The relevance of a document is calculated (using Bayesian statistics and log-odds) as the summation of probabilities of terms which co-occur in the document and the query:

$$P(Q, D_i) = \sum_{t_i \in Q, D_i} \log \frac{p_i(1 - q_i)}{q_i(1 - p_i)} \tag{7}$$

where

p_i = Probability that a document contains term t_i given that it is relevant, $P(t_i|R)$
q_i = Probability that a document contains term t_i given that it is non-relevant, $P(t_i|\overline{R})$

The appropriate substitutions for p and q are the proportions:

$$p = \frac{r_i}{R} \tag{8}$$

$$q = \frac{n_i - r_i}{N - R} \tag{9}$$

where

N = Number of documents in the collection
n_i = Number of documents in which term i occurs
R = Number of known relevant documents in the collection
r_i = Number of known relevant documents in which term i occurs

Substituting the values of Equations (8) and (9) into Equation (7) we obtain the relevance weighting formula of Robertson and Sparck Jones (1976):

$$w_i = \log \frac{(\frac{r_i}{R})(1 - \frac{n_i - r_i}{N - R})}{(\frac{n_i - r_i}{N - R})(1 - \frac{r_i}{R})} \tag{10}$$

As it is usually not possible to know the number of relevant documents in the collection for a given query, R, estimation for the values of p and q must be made. This

can be done by taking a sample document and query collection and retrieving relevance judgments on this sample set. This is however not always possible and so Croft and Harper (1979) proposed a different approach which assumes p_i (the probability a document contains term t_i given it is relevant) is the same for all query terms and $\frac{p_i}{1-p_i}$ is constant and can be ignored for ranking purposes.

The most commonly used probability model implementation is the BM-25[5] model introduced by Robertson et al. (1994) at the third TREC conference in 1994. It was a combination of previous models used by the City University team and aimed to incorporate the document length into calculations of relevance. Equation (11) also uses weighting functions the team had introduced the previous year to incorporate term frequencies.

$$BM25(q,d) = \sum_{t_i \in Q} \log \frac{(r_i + 0.5)(N - n_i - R + r_i + 0.5)}{(R - r_i + 0.5)(n_i - r_i + 0.5)} \times \frac{(k_1 + 1)tf_i}{K + tf_i} \times \frac{(k_3 + 1)qtf_i}{k_3 \times qtf_i} \tag{11}$$

where

$$K = k_1((1 - b) + b \times dl/avdl)$$

N = Number of documents in the collection
n_i = Number of documents in which term i occurs
R = Number of known relevant documents in the collection
r_i = Number of known relevant documents in which term i occurs
$tf_{i,j}$ = Term frequency measure of term i in document j
$qtf_{i,j}$ = Term frequency measure of term i in the query
k_1 = Constant which determines the influence of $tf_{i,j}$
b = Constant which determines the influence of document length normalisation
dl = Document length of document d
$avdl$ = Average length of documents in the corpus

For a typical retrieval task of retrieving a list of results in response to a user specified query and ignoring any repetition of terms in the query, as is the case for the vast majority of web queries, Equation (11) can be simplified to:

$$bm25(q,d) = \sum_{t \epsilon q} log\left(\frac{N - df_i + 0.5}{df_i + 0.5}\right) \times \frac{(k_1 + 1)tf_i}{k_1((1 - b) + b\frac{dl}{avdl}) + tf_i} \tag{12}$$

where df_i is the number of documents in the collection that contain the term i.

[5]BM = Best Match

2.1.3 Linkage Analysis

The study of the linkages between documents is not a new area of research, however it has found increased popularity due to the hyper-linked structure of the World Wide Web. Linkage analysis is concerned with the links made between entities within a collection. Early forms of this were the science of Bibliometrics and Citation Analysis as discussed in Chapter 4. The links between scientific papers and articles was seen as a way to measure the influence and importance of scientists within the scientific community (Garfield, 1972). Linkage is seen as a good indicator of human judgments on the value of a document or information source; by linking to another document (in the case of bibliometrics this may be another paper, on the web another web page) the creator of the link as provided an explicit judgment of authority and similarity between the two sources (Chakrabarti et al., 1998).

Linkage analysis provides documents with a measure of importance based on the network of documents which connect to them; news forum search and e-mail retrieval both benefit from this approach. Using linkage analysis it is possible to retrieve documents which are highly connected; in the context of e-mail search, say, this means it is possible to only consider e-mails which have received some minimum number of replies.

Linkage analysis techniques form the basis of one of the best known search engine providers in the world. As with most link analysis techniques, the techniques used are of an iterative nature, allowing the importance attributed to each web page by the 'in-links' (connections made to a web page) to propagate to other web pages. We shall now discuss two of the most famous approaches proposed for using linkage information to aid in relevant document retrieval, *PageRank* and *HITS*.

2.1.3.1 PageRank

PageRank is one of the best known linkage analysis techniques and forms the basis of the Google search engine (Google Inc., 2006). It is a query-independent retrieval strategy which takes the form of a random-walk (Motwani and Raghavan, 1995) by a web user over a web-graph. The user assumes the role of a *random surfer* who randomly chooses a web page. From here, the surfer clicks on random links within the page, following the links to another page and never clicking 'back'. Eventually the surfer becomes bored and selects a new web page at random and begins surfing again. This boredom is modeled in the PageRank (Page et al., 1998) algorithm by the inclusion of a *'dampening factor'*.

The PageRank of a document, d, is the combined PageRank of every document in the set, S, of documents which link to d (*indegree*) divided by the number of outlinks (*outdegree*) from each document in S. The PageRank score of a document is achieved

through convergence of an iterative algorithm.

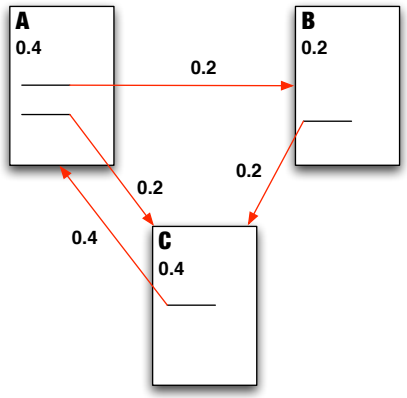


Figure 6: Simplified PageRank calculation (Page et al., 1998).

To calculate PageRank, an initial PageRank score PR_n is assigned to each web page. We then calculate a simple PageRank score for each document as follows:

$$PR'_u = c \cdot \sum_{v \epsilon S_u} \frac{PR_v}{outdegree_v} \tag{13}$$

where c is a constant that is maximised and < 1, S_u is the set of documents that link into document u.

The value PR'_u is calculated iteratively until a suitable convergence is achieved. Page et al. (1998) report acceptable convergence ranks in 52 iterations for a crawl of 322 million links, while convergence on half that data takes roughly 45 iterations. Under certain circumstances however, this simple PageRank formula is susceptible to certain problems which do not allow their scores to be propagated back into the rest of the linkage graph; *dangling links* can be created when a page has not been downloaded but a link points to this page; *rank sinks* exist between pages which point to each other, but do not point to anything else creating a loop or trap so accumulated scores are never distributed.

To overcome these problems, a *rank source* vector $\overrightarrow{E}$ may be introduced which has in-links from all other nodes in the web graph. This ensures that the iterative scores are distributed back into the graph, as illustrated in Figure 7. This also means that a web surfer is never confined to following a specific path through the graph and is always able to become 'bored' and jump to a different location. The score of $\overrightarrow{E}$ itself is usually distributed uniformly across all the other nodes of the graph, however it is possible to create a more personalised variant of PageRank by changing the distribution of this vector.

The weight accumulated in the $\overrightarrow{E}$ vector is usually distributed equally across all nodes in the graph (in most experiments Page et al. (1998) used a uniform vector

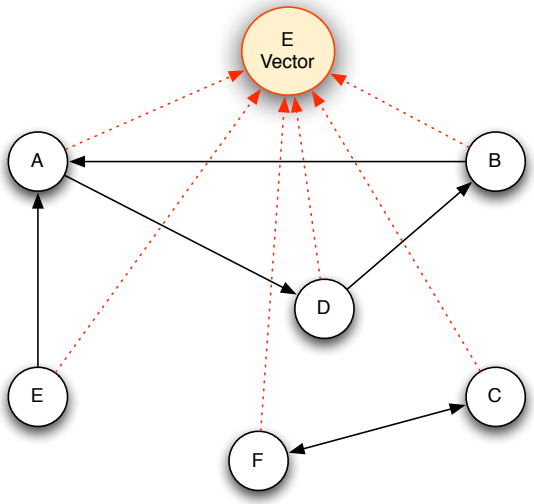


Figure 7: The introduction of an E vector

with $||E||_1 = 0.15$), however by tailoring the distribution of weights we can create a personalised PageRank, based on a user's preferences. This simulates a change in the browsing behaviour of the "random surfer", making it more like a specific user, or enables a more topic-specific style of PageRank (Page et al., 1998; Haveliwala, 2002). The new PageRank which includes the $\overrightarrow{E}$ vector is calculated as follows:

$$PR'_u = c \cdot \sum_{v \epsilon S_u} \frac{PR_v}{outdegree_v} + c(\overrightarrow{E}(v)) \tag{14}$$

where $\overrightarrow{E}(v)$ is the value of the $\overrightarrow{E}$ vector that is to be be distributed back to document v.

PageRank is calculated independent of any query and so does not affect the query execution time. When combining PageRank with a query-specific scoring mechanism, care must be taken so as not to introduce *topic drift* or *topic distillation* (Chakrabarti et al., 2001; Bharat and Henzinger, 1998). This can occur when pages with high PageRank are highly ranked even though they are not relevant to a user's query.

2.1.3.2 Hyperlink Induced Topic Search

Unlike PageRank, Hyperlink Induced Topic Search (HITS), proposed by Kleinberg (1998) is a query-dependent form of linkage analysis. The algorithm is a two-stage process; a query is first issued to a standard search-engine, returning a subset of documents. Two *mutually reinforcing* scores are then calculated for each of the documents in this subset:

Authority: A good authority page is one which links to many other pages which are relevant to the query, whilst containing (a large amount of) information relevant to the query itself.

Hub: A good hub page is one that can "pull together" authoritative pages by containing many links to authority pages.

In order to calculate the two scores for each page within the initial *base set* we perform the following steps (suggested numbers are taken from (Kleinberg, 1998)):

1. Retrieve an initial set (top 200 say) of relevant documents (referred to as a *base set*).
2. Expand this set by following off-site inlinks as well as off-site outlinks to produce an expanded set of relevant documents (1,000 - 5,000 documents).

An iterative algorithm makes use of the mutually reinforcing nature between hubs and authorities, maintaining and updating the numerical weights for each page; for each page p, a non-negative authority weight ($x^{\langle p \rangle}$) and a non-negative hub weight ($y^{\langle p \rangle}$) is calculated. After each update, the weights of each type are normalised so their squares sum to 1 so as to remain invariant:

$$\sum_{p \in S_\sigma} (x^{\langle p \rangle})^2 = 1 \tag{15}$$

and

$$\sum_{p \in S_\sigma} (y^{\langle p \rangle})^2 = 1. \tag{16}$$

Pages with larger x and y values are viewed as being "better" authorities and hubs, respectively (Kleinberg, 1998).

The HITS approach aims to tackle the *abundance problem, the number of pages that could reasonably be returned as relevant being far too large for a human user to digest.* It suffers from a few drawbacks though, such as poor selection of the base set. If the initial query does not cover a sufficiently broad topic, there will often not be enough relevant pages in the base set from which to extract a sufficiently dense sub-graph of relevant hubs and authorities. The main disadvantage however, is that the two scores are calculated at query time which requires extra resources from the search system at query time, but also increases the system response time. This represents a major disadvantage to the general user, who requires the minimum delay in system response.

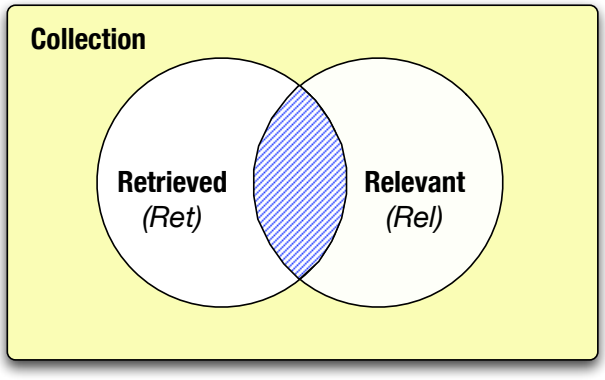
2.1.4 Evaluating Retrieval Performance

The context in which a retrieval system is to be used plays an important role in deciding it's performance and how it performs in an evaluation. As mentioned in Section 2.1.1.2, the vast majority of the time users will not look beyond the first page of results to find a relevant document. This means that although there may be large numbers of relevant documents for a query, users will only ever see the first 10 or perhaps 20. (Google for example displays just 10 results per page.) While this is sufficient for finding general information about, say a holiday, in a more specific and exact search (e.g. a doctor's search for an exact match to symptoms) a user may need to have all relevant information before making a decision.

2.1.4.1 Precision and Recall

Precision and recall allow us to measure the amount of available relevant information we currently have, versus the amount of relevant information available to the retrieval system. Consider a query made to a retrieval system. In response to this query, the system returns the set of documents, *Ret*, which it believes are relevant to the query. This may or may not contain some or all of the set of relevant documents, *Rel*, available to the system. This is illustrated in Figure 8.

Two complimentary measures are often used to measure a retrieval system's performance and they are defined below. **Precision** is the fraction of the documents found within a certain cut-off point which are relevant. **Recall** is the fraction of the total relevant documents found by the system within a certain cut-off point.



$$Precision = \frac{|Rel \cap Ret|}{|Ret|}$$

$$Recall = \frac{|Rel \cap Ret|}{|Rel|}$$

Figure 8: Retrieved documents vs. relevant documents

The ultimate goal of any retrieval system is to obtain high precision and high recall. This is however a very difficult task as it is accepted that the two have an inverse relationship; the higher the recall, the lower the precision (Figure 9). It is therefore a

more realistic goal to tailor the system to the needs of the user, finding a suitable balance between these two measures. Taking as an example the situation given at the start of this section; in the case of the general web user searching for holiday information, this requires high precision so as to get as much relevant information with as little effort as possible; in the case of the doctor searching for treatments, high recall is vital so as to retrieval all relevant information.

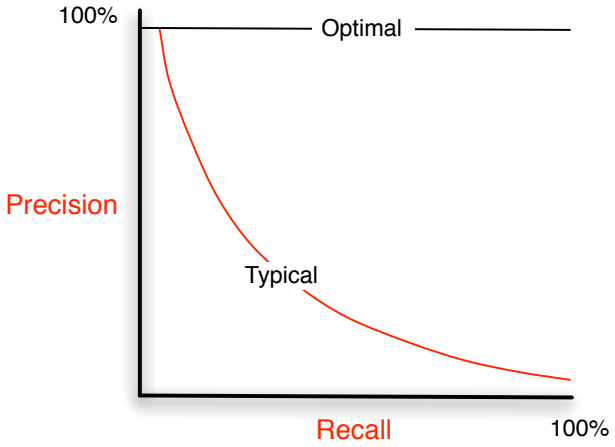


Figure 9: The precision-recall curve

2.1.4.2 Single Value Performance Measures

While Precision and Recall provide measures of a system/algorithm performance, it can be desirable to indicate this performance using just a single figure. The following measures indirectly combine both precision and recall into a single measure.

Average Precision(AP) provides a measure the precision for a query at each point where a relevant document is found:

$$AP = \frac{\sum_{i=0}^{N} P(i)}{N} \tag{17}$$

where:

N is the total number of relevant documents
$P(i)$ is the precision at document i.

Mean Average Precision (MAP) provides a measure the overall precision of a system by averaging the average precision over all queries made by the system:

$$MAP = \frac{\sum_{j=0}^{Q} AP_j}{Q} \tag{18}$$

where:

Q is the total number of queries
AP_i is the average precision for query j.

2.1.5 Combining Sources of Evidence

It is a statistical and an intuitive fact that the more evidence that can be provided to support an hypothesis, the more likely it is that the hypothesis is true (Croft, 2000). This fact is strong motivation for combining the sets of documents retrieved by different search systems in response to a query, since it has been shown that there is surprisingly little overlap in the sets retrieved by different search systems (Harman, 1993). This can be the result of many different factors; document representation can play a large part in the documents retrieved. Different representations of a document (using title and abstract versus free-text and manually assigned index terms) can be combined to aid in the retrieval process (Croft and Harper, 1979; Das-Gupta and Katzer, 1983). The way in which algorithms treat a query can also affect the retrieved set, as shown by McGill et al. (1979). Das-Gupta and Katzer (1983) found that while the overlap in retrieved document sets can be very low, the overall performance in terms of recall and precision (see Section 2.1.4.1) remained very similar. They also demonstrated (in confirmation of the findings of McGill et al. (1979)) that searcher tendencies, and the way in which different searchers approach a retrieval problem, can not account for the low levels of overlap alone.

By combining the outputs from several different retrieval methods, the overall ranking of relevant documents can be improved. The problem still remains however of how best to combine these outputs for optimal performance. The increased recall gained by combining multiple retrieved sets can result in a decrease in precision due to the inverse relationship of the two (Cleverdon, 1972). Each source of retrieved documents may however be seen as a further expert opinion on the documents retrieved (Bartell et al., 1994), providing further evidence of their relevance. The aim of combination is to reduce the errors which can be made in ranking documents. Fox and Shaw (1995) state that there are two major errors which can be made in ranking (akin to type-I and type-II errors in statistics): ranking non-relevant documents highly and ranking relevant documents lowly. One important consideration here is that the scores given to documents by different ranking algorithms and systems may be very different and incompatible. If, for example, one scoring scheme attributes scores in the range [0, 1] and another in the range [-100, 100], then the effects of combing these two scoring schemes

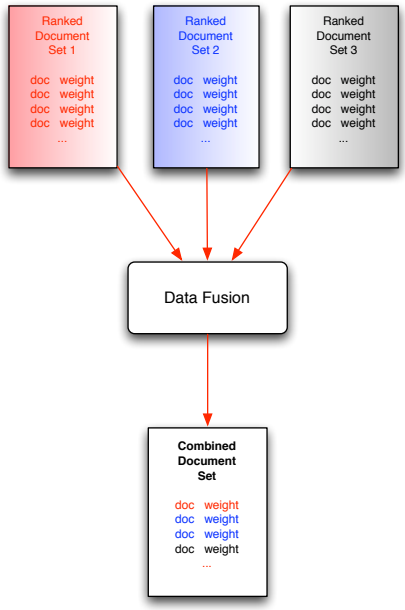
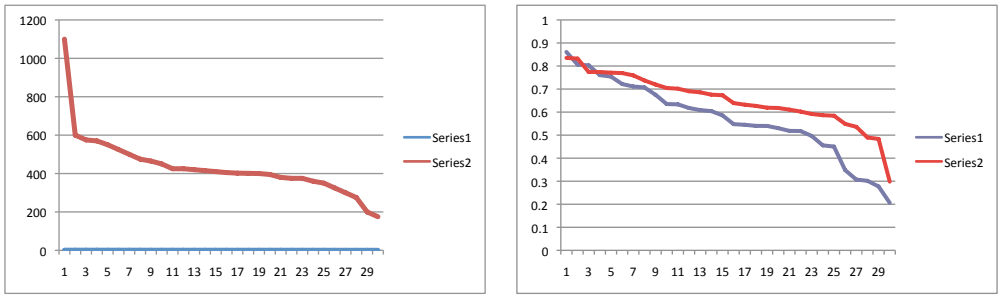


Figure 10: Combination of different result lists.



(a) Unnormalised Sources (b) Normalised Sources

Figure 11: Combination of non-homogeneous sources

based on scores alone will be very small; the second score will have a disproportionate affect on the combined ranking. This is shown in Figure 11(a). These scores must be in some way be *normalised* so as to allow an even distribution of the effects of each scoring algorithm. The scores in Figure 11(b) have been normalised using the commonly used *min-max normalisation* technique.

$$s'_i = \frac{s_i - min\{s_i\}}{max\{s_i\} - min\{s_i\}}(new\_max - new\_min) + new\_min \tag{19}$$

This has the effect of leaving all scores, s'_i, within the range new_max to new_min (commonly 0 to 1), and removing any over-bearing influence of a single ranking scheme s_i.

Two two major approaches to combining multiple sources of evidence are now discussed, namely similarity merger, and linear combination.

2.1.5.1 Similarity Merge

Several combination or *data fusion* techniques were proposed by Fox and Shaw (1995) based on the unweighted min, max or sum of each document's normalised score. The two most successful of these were CombSUM and CombMNZ, which calculate a combined score for a document d, from a number of data sources as follows:

CombSUM:

$$Score(d) = \sum_{i=0}^{n} Score_i(d) \tag{20}$$

where n is the number of data sources that are to be combined.

CombMNZ:

$$Score(d) = (\sum_{i=0}^{n} Score_i(d)) \times k \tag{21}$$

where k is the number of times $Score_i(d) > 0$

The six different approaches proposed by Fox and Shaw (1995) combine the similarity scores assigned to documents by ranking procedures. Lee (1997) later found that the effect of combining the actual ranks assigned to documents was not as effective, except in the case where the search systems had very different characteristics in terms of the shape of the score-rank curve. This can be interpreted as evidence that the normalized score is usually a better estimator for the probability of relevance than the rank. It should be noted however, that the CombMNZ method may penalise documents that do not occur in one or more of the result lists when applying the fusion method to the *top-n* documents. This is a problem since it has been observed that although a source may provide a poor ranking on its own, it may aid relevance judgement more effectively as part of a combination (Bartell et al., 1994).

2.1.5.2 Linear Combination

Similarity merge techniques are used extensively in IR systems to combine the output of several retrieval sources. While Fox and Shaw (1995) use an equal weighting for each of the sources to be fused, they note that weighting sources which perform more strongly is a consideration. Bartell et al. (1994) and Vogt and Cottrell (1999) have both made this same suggestion so as to not disregard a poor ranking scheme, whilst also ensuring

that the combined ranking is as optimal as possible. This approach is referred to as *Linear Combination*, and is calculated as follows:

$$Score(d) = \sum_{i=0}^{n} Score_i(d) \times w_i \tag{22}$$

where w_i is the weight associated with the data source i.

The calculation of weights can be done in a number of different ways; Thompson's 'Combination of Expert Opinion' model gives weights to each source based on the past performance of the source system; Bartell et al. (1994) choose to assign weights based on a training phase performed using a set of training queries.

Now that the basic measures and methods have been discussed, we shall look at the situations in which these methods will be used within the context of this thesis. Social network analysis will play a key role in determining the graphs and networks across which linkage analysis and information retrieval shall be implemented. We discuss the main ideas and approaches within the field below.

2.2 Social Network Analysis

Social network analysis is the study of social relationships between individuals in a society. The focus of the analysis is not on the attributes and qualities of the actors involved, but rather on the ties which link them. A tie exists directly between two actors, although groups of actors may be related through some common goal or concept (Wasserman and Faust, 1994).

Social network analysis has been studied in connection with a wide range of areas including information retrieval (Yu and Singh, 2003; Zhang and Ackerman, 2005) and trust (Windley et al., 2007). The most relevant part of social network analysis to this thesis is that of the "Small World" literature. It is from this literature that we get the commonly heard phrase "Six Degrees of Separation" (Guare, 1990) which originates in the work of Stanley Milgram. He and a co-worker, based their ideas on those of Pool and Kochen (1978) which although finally published in 1978 had been circulating for nearly two decades prior to that. Pool and Kochen had been interested in the mobilisation of political power through the contacts made by politicians. They had suggested that one of the informal ways in which associations and alliances were created was at cocktail parties. This then led to the question "what is the probability that two strangers will have a mutual friend?".

2.2.1 The Small World

Milgram advanced this idea in his seminal work "The Small World Problem" (Milgram, 1967). Milgram distributed letters addressed to a stockbroker friend of his, to strangers in Nebraska. Each letter came with the instruction to pass the letter to friends (those that were known on a first-name basis) with the ultimate goal of the letter reaching a particular stockbroker in Boston. Later similar experiments found that geographic proximity and similarity of profession to the target person were the most frequently used criteria by subjects for selecting a friend to pass the letter to (Dodds et al., 2003). Based on the number of people through whom the letters traveled, Milgram concluded that everyone in the country was connected through a chain of at most six people. There have since been questions raised about both the scientific rigour of both the experiment and the conclusions drown by Milgram (Kleinfeld, 2002). Subsequent experiments (Korte and Milgram, 1970) have demonstrated that two randomly chosen people are in general connected by only a few intermediate connections, and there is widespread acceptance of the initial results.

While the study of the social connections between people in society may at first seem unworthy of serious scientific research, it is important to consider that it is through these networks that the vast amount of human knowledge actually flows. As mentioned in Section 2.3, a lot of the information we get comes directly from friends, work-colleagues and other direct acquaintances. The power of this network to transfer information through local contact to the global network cannot be underestimated. More importantly, it has been shown that the networks formed by people is only one example of a network which allows information (or any form of message) to spread quickly, others being the internet (Jeong et al., 1999), power grids (Watts and Strogatz, 1998) and the spread of disease. All these examples exhibit similar qualities to the social networks formed by people. It is possible therefore that the development of effective models of social networks will improve our understanding of many other fields as well.

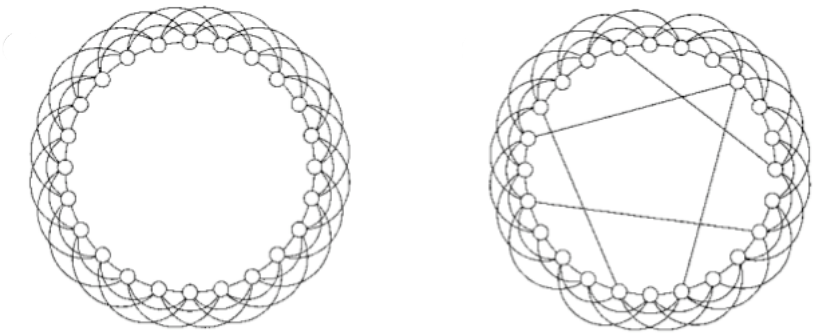
On a lighter note, people have examined the networks that arise through social collaboration such as co-starring roles between actors; Brett Tjaden's parlor game "The Six Degrees of Kevin Bacon" (Tjaden and Wasson, 1997) and the network of collaboration between authors within a particular conference series (Smeaton et al., 2003) have both been studied. These studies are based on previous examinations of the co-authorship network created around the highly-respected and prolific mathematician Paul Erdös and the so-called "Erdös numbers". (It is somewhat fitting that this should be the case since Erdös is one of the fathers of random graph analysis, research which social network analysis builds upon.) Erdös was an Hungarian mathematician who traveled Europe and the US extensively, collaborating with a vast number of fellow mathematicians. In his book "The Man Who Loved Only Numbers" (Hoffman, 1998) tells the extraordinary story of the man who would effectively pay for his keep while staying with friends by co-authoring

papers with them. An "Erdös number" is the smallest number of co-authorship links between an individual and Paul Erdös. An extensive website[6] exists which allows people to look up their own Erdös number and has much more related information[7]. The concept of Erdös numbers still captures the imagination of many researchers.

2.2.2 Social Network Models

Since Milgram's paper was published, the concept of small worlds has been formalised mathematically. Small worlds are characterised by two main properties. Firstly, the average path length between any two nodes in the graph grows logarithmically with the size of the graph. Random graphs are the simplest incarnation of a small world, and have been extensively studied in the past, particularly by Erdös and Rényi (1959). Random graphs however do not exhibit the second property required of small worlds; social networks have a high degree of *connectivity* compared with random graphs.

The connectivity or *clustering coefficient*, C of a graph is a measure of the fraction of connections between neighbours of a node, n, that actually exist compared to the total number of possible connections. In a fully connected network, in which everyone knows everyone else, $C = 1$; in a random graph $C = \frac{z}{N}$, (where z is the average number of connections between nodes) which is very small for a large network. In real-world networks it has been found that, while C is significantly less than 1, it is much greater than $O(N^{-1})$ (Watts and Strogatz, 1998; Newman, 2000).



(a) (b)

Figure 12: 12(a) Friends-of-friends become friends 12(b) Long-distance friends included

One of the first and most widely studied models of small worlds was proposed by Watts and Strogatz (1998). Their model attempts to make up for the shortcomings of the random graph. A random graph is created by taking a bunch of nodes and connecting

[6] http://www.oakland.edu/enp/

[7] The author of this thesis has an Erdös number of 5 as he has co-authored with Alan Smeaton, who has co-authored with Nicola Stokes, who has co-authored with Alistair Moffat, whose PhD supervisor and co-author has himself co-authored with Erdös

randomly chosen pairs together with lines or edges. (A more formal overview of graph theory is given in Section 4.4.) In general, the connections made by people within a community is not created solely at random. People make new acquaintances through current acquaintances, with friends-of-friends becoming our own friends. This idea is shown in Figure 12(a). Each node is connected not only to its own neighbours, but to its neighbour's neighbours. Increases in the level of connectivity are discussed in (Watts and Strogatz, 1998; Newman, 2000). As well as this, Watts and Strogatz (1998) randomly "rewire" the links between each node with probability p, keeping the average number of connection constant, but increasing the clustering coefficient. This step introduces a far more realistic element to the network; as well as being introduced to people by our local friends (friends who live in the same neighbourhood or work with us), we retain friendships with people we have met who are a long distance from us. In a social sense, this may be people who live far away from us, or acquaintances from previous eras of our lives. In the same paper, Watts and Strogatz (1998) shows that this model is applicable to the network created by the neural network of the worm *Caenorhabditis elegans*, the power grid of the western United States, and the collaboration graph of movie actors.

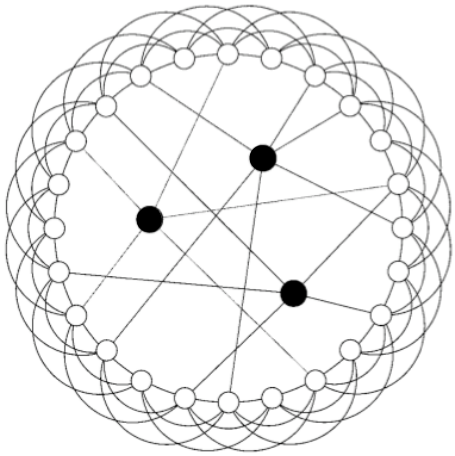


Figure 13: Small world model of Watts and Storgatz with additional "connectors"

Kasturirangan (1999) proposed a different model which accounted for the small world phenomenon present in social networks. This phenomenon was due not to the presence of long-distance connections between nodes, but instead because of a few very highly-connected nodes. These nodes are shown in black in Figure 13 and represent people who exhibit a high degree of connectivity. Through them, the short average path lengths are achieved. The Episcopalian minister in Milgram's original experiment is a good example of this. In his book "The Tipping Point", Gladwell (2000) refers to these people as "connectors".

In our work we shall be looking at the co-authorship and citation networks of authors within the SIGIR community, a community which also exhibits the small world

characteristics discussed here. We will now look at the trust that people put into the information which they receive from the network around them because, as we show later, trust is also incorporated into our work.

2.3 Trust and Authority

Trust is a complex notion involving many different considerations. It has in the past been viewed in the context of recommender systems (O'Donovan and Smyth, 2005), economics (Das-Gupta, 1988), online interactions (Friedman et al., 2000; Van House, 2002), social networks (Golbeck and Hendler, 2004) and information retrieval (Briggs and Smyth, 2007). Trust was originally studied in the context of encryption and security but there has since been a growth in research which views trust from a more interactive and societal standpoint. With the growth of the internet and the ease of publication of information, there has been a corresponding growth in the number of different sources of information. These sources tend to have just sprung up without any track record or history and so have no historical authority. As a result, there is increased demand for novel and improved means of validating new information sources. By validation we refer not only to the information contained within, say blogs and wiki-style publications, but also the sources of the information themselves(Guha et al., 2004; Rieh and Belkin, 1998).

Trust plays a key role in our everyday lives, from trusting a shop to not overcharge a credit-card transaction, to trusting the credit-card company itself with storing the information securely. The trust we place in these institutions as well as the people we interact with in general is just one aspect of trust. These trusting assumptions that we make are based on a variety of factors, a major part of which is past experience on our part, as well as the experiences of the people we know. These are two different types of trust; the trust in our own experiences, and the trust we place in people we know to tell us the truth.

When we place our trust in an entity (be that a person or an organisation) we do so in the hope that they will not betray or abuse that trust. It is not merely this that guides our decisions of whom to place our trust in. Das-Gupta (1988) makes the point that our decision is also guided by the fact that *"knowing what you know of his [the trusted party's] disposition, his available options and their consequences, his ability and so forth, you expect that he will choose to do it."* Trust is based on a conscious acknowledgement and assessment of risk, and this is what differentiates it from other related ideas such as faith (Chopra et al., 2003).

Trust and Authority are two concepts which help to bring organisation to a system. That system may be the network of acquaintances of a person, the business dealings

of an organisation, or the interactions between information systems. Without some understanding of the other agents or entities within a system, all the actions we take are taken without prior knowledge of past events, events which could influence the outcome of a decision on whether interaction should take place or not.

2.3.1 A Trust Framework

While trust has been defined in many ways, there is an important commonality which needs to be stated in all of these definitions; trust is needed only in situations where there is incomplete situational knowledge on the part of the party conferring trust (the trustor) on another (the trustee) (Abdul-Rahman and Hailes, 2000). If we have perfect knowledge of a proposed transaction, there is no need for trust. The lack of vital knowledge by one side however, can be fatal to an arrangement of trust. This was famously investigated by Akerlof (1970) who showed that in a market where one side has perfect knowledge (the seller say) and the other not, a complete breakdown of trust and therefore of trade can arise. Trust is used to reduce the complexity of a situation; by conferring trust on a party, we remove the necessary creation of judgments and observations since we shall be trusting someone to do or have done these for us.

The types of interactions which are based on differing degrees of trust have now found their way into the virtual world, allowing for e-commerce to flourish, as well as the growth of social networks and social computing. One of the issues with online interaction is the lack of what Axelrod (1984) refers to as the *"shadow of trust"*; without a sufficient deterrent to prevent a party from breaking a trust agreement, there is no reason for them not to do so. This deterrent comes in the form of a history, a memory of past interactions. For example, a shop which is known to trade in counterfeit products will receive less trade since people will not trust it. Axelrod (1984) himself gives the far better example of the co-operation that was engendered during the First World War between opposition fighters in opposing trenches; cessation of fighting occurred due to an uneasy truce caused by either side deciding, as Das-Gupta (1988) stated, that the consequences of breaking the truce (i.e. renewed fighting) made a truce the best option.

Trust itself has, in the past, been divided up into two main areas; *cognitive* and *emotive* (or affective) trust (Craig, 2008). Emotive trust deals with trust which is based on emotion, it is normally concerned with situations where there is some sort of bond between the two parties involved. An example of this is trust between family members, or the trust that exists through people who may identify with others of the same philosophy.

Cognitive trust is a more considered approach, relying on risk assessment and consideration of past experience. This is the type of trust found in online environments

where trustor and trustee may not necessarily ever meet face to face. It covers such assumptions as reliance on the other party to do as they have said, as well as their competence in doing so. The way in which a person will act based on an assumption of trust made about another person or institution is also part of cognitive trust.

Van House (2002) discusses cognitive trust in terms of *cognitive authority*. She quotes Wilson (1983) who provides the useful distinction between cognitive authority and expertise; a person can be a known expert in a field, but we grant cognitive authority to anyone we ask advice of. Expertise may be seen as a globally judged measure of trust in the opinion or action of an individual/institution based on the observations of society as a whole. Cognitive authority is ascribed to any person/institution by a single person in a particular context.

Chopra et al. (2003) gives an overview of trust based on an extensive literary review. The four categories of trust mentioned in his work are similar to those of Abdul-Rahman and Hailes (2000):

- *Interpersonal Trust* as stated by both papers, is the trust that the trustor places in the trustee directly. This trust is specific to both trustee and context. For example, while I may always trust the opinion of my astrologer friend in astral matters, I may not trust their restaurant recommendations.
- *Dispositional Trust* or *Individual Trust* is a form of emotive trust; I trust that in acting in a certain way towards others, they will treat me accordingly. As it is independent of both context and the parties involved, it may be thought of as a naive trust.
- *Societal Trust* or *Systemic Trust* refers to the trust that is placed not in any specific agent or institution but more the rules that govern a system of interaction. The monetary system or rules of the physical world are examples; we trust that money is worth a certain amount when accepting it as payment, and we continue to trust that apples will fall from trees.
- *Relational Trust* is only presented by Chopra et al. (2003) and may be seen as part of societal trust, but is worthy of separate mention. This is the trust that springs from recurrent interaction with the same trustee and arises as a consequence of this interaction. Chopra et al. (2003) cites Seligman (1997) in describing this trust as the "social glue" which holds society together.

It is generally agreed that trust is a social and psychological phenomenon, though philosophical interpretations have been made (Hirschman, 1984). In order for society to function smoothly and continuously, we are necessitated to make several trustful assumptions every day. Van House (2002) points out that the cost (maybe not monetarily

but in terms of time and effort) involved with obtaining perfect knowledge of a situation, thereby removing the need for trust, is prohibitively if not impossibly high. *"We have neither the ability nor the resources to make all possible observations, develop our own methods, and test all possible claims"* (Van House, 2002). A result of this is that we must trust in others, in their observations, and the communication of those observations that they make to us. How and why we decide to trust others is discussed in the following section.

2.3.2 Reputation

> *"Reputation, reputation, reputation! Oh, I have lost my reputation! I have lost the immortal part of myself, and what remains is bestial."(Cassio; Othello. ACT II Scene 3.)*

The role of reputation in the Shakespearian play Othello is great. Loss of reputation and a desire to regain it drives the play forward to its tragic conclusion, but what is reputation? We shall discuss this below, but from a simple standpoint it is necessary to note that the perception and expectation of a person's actions can have profound effects on the levels of trust placed in that person. Friedman et al. (2000) cite the excellent and well reported stories of two online companies Trustee (NYTimes, 1999) and Amazon (Rosman, 1999) (see also Economist (2001)). Both of these companies were thought to have abused the trust of their public, either directly or indirectly through laxity. Reputation has proven time and again to be something which is hard to earn, and easy to lose.

The "shadow of trust" (Axelrod, 1984) is related to the field of game theory which has seen much research into trust (Kreps and Wilson, 1982; Dellarocas, 2003). The deterrent of a *"shadow of the past"* is investigated by Friedman and Resnick (2001) who showed that removing this deterrent is detrimental to the system as a whole. The idea of this shadow is to show to the rest of the system/population that the agent/party is worthy of trust. In order to do this however, the 'shadow' must be created. The creation of this shadow is done through interaction with various parties, or simply the same party if we are to create a one-to-one bond between parties. This shadow then equates to the idea of a reputation which has been put forward by Abdul-Rahman and Hailes (2000):

> *"A reputation is an expectation about an agent's behaviour based on information about or observations of its past behaviour."*

Hirschman (1984) favours the ideas put forward by Arrow (1962) that trust is amongst the *"resources whose supply may well increase rather than decrease through use; second,*

these resources do not remain intact if they stay unused; like the ability to speak a foreign language or to play the piano, these moral resources are likely to become depleted and to atrophy if not used". Since resources tend to deplete with use, he concludes that trust is in some ways a skill or quality rather than a resource. The idea that trust grows with use gives rise to the creation of a reputation; reputation may also fade if no trusting interactions are made for a long period of time.

In discussing prerequisites to trust, Chopra et al. (2003) reason that reputation is the act of placing trust in trust. We infer the trust that we have in a trustee based on the trust that others have placed in that trustee. While there are many reasons for placing our trust in a person/institution (such as a bond, or identification of a similar goal), if we do not have enough information about them through our own interactions and history then we must rely on others. Trust must be propagated and distributed through society so that reputations (both good and bad) may be used to improve the judgments and overall quality of agents' interactions. In the next section we shall show how this propagation takes places, taking as a focus the propagation of trust in online environments.

2.3.3 Propagation of Trust

The way in which trust and reputation is distributed across a network of agents is strongly related to the social networks of Section 2.2. The most common method of spreading or creating a reputation is by *word-of-mouth*; after interacting with a trustworthy party, we will most likely tell other people that we know about the trustworthiness of that party. This is in fact the most common way in which people decide upon using a new brand or service (Das-Gupta, 1988). Information we get from our friends and colleagues about products is considered more trustworthy than information we get from a random source such as an advertisement. This is because of the history we have already created with our friends and colleagues, as well as other factors as discussed in Section 2.3.1.

The interaction that people have with computers and the internet are taken very personally. It has been shown that the trust placed in, say, search results is exactly the same as the trust that would be placed in a person. Also the trust which is placed in hardware itself; a person may act as though a trust has been abused when a machine breaks down (Chopra et al., 2003). This is somewhat at odds with the intuitive idea put forward by Friedman et al. (2000).

2.3.3.1 Online Trust

In an online environment, a lot of the checks which we would normally make into the trustworthiness of another party are not available. The growth of the web has brought with it the possibility for enormous numbers of people to learn of opportunities and interactions that would not have been possible without it: *"Already, internet-based reputation systems perform commercial alchemy. On auction sites, for example, they enable trash to be shuttled across the country and in the process transmuted into treasures."* (Resnick et al., 2000). An important message lies within this statement however; the opportunity to interact with others brings an increased opportunity to shuttle 'trash' and lemons around (Akerlof, 1970). As Das-Gupta (1988) points out, without the appropriate mechanisms for penalising disruptive or distrustful behaviour, individuals will not possess the appropriate incentives to act truthfully; and since this will be generally recognised within the population, people will not wish to enter into transactions with one another.

E-commerce has seen rapid growth since its beginnings despite the problems mentioned. This is due to a number of innovations in the field of reputation creation. One method of helping to both weed out the dishonest agents and allow others to build a proper reputation is to provide a simple feedback mechanism to users. The system employed by eBay has been extensively studied (Houser and Wooders, 2006; Resnick and Zeckhauser, 2002). Dellarocas (2003) gives a good overview of the different aspects of the eBay feedback mechanism which have been studied. This approach relies on the 'quantity over quality' idea; although we may not know the agents (or indeed their own reputation rating) providing the ratings of an agent, a sufficient volume of ratings will help to determine trust. By introducing the ability to provide feedback on those that we interact with, a history may be built and a 'shadow' created. Interestingly however, it has been found that there are several weaknesses with the feedback approach. Resnick and Zeckhauser (2002) found that a surprisingly high percentage of comments were positive. One of the causes of this is an apparent culture within eBay users to negotiate before posting negative feedback. There is also the fear of retaliatory negative feedback which is akin to the 'mud-slinging' of political campaigns.

The reputation built by the eBay feedback model is one which relies on a global trust value; the information used in creating a reputation is taken from many different and disparate sources, none of whom may know each other directly. This is very different to the word-of-mouth model in which information is passed on a local level; I learn what I know about others through direct interaction, or through the interactions of my close acquaintances. The global model of trust has been criticised for its lack of both context and personalisation; trust and reputation are most often context-sensitive, and trust specifically is a personal quality.

There has been a great deal of research into the creation of metrics which are more firmly grounded in the aspects of trust discussed in Section 2.3.1. Sabater and Sierra (2005) gives an extensive overview of the most highly cited interpretations of trust in a computational and online environment. The first to propose a general computational model of trust was (Marsh, 1994). He proposed a highly complex model which took into account many of the factors influencing trust as discussed by Chopra et al. (2003). The complexity of Marsh's model has been criticised for introducing large numbers of variables, these variables being used to model concepts such as 'risk' and 'competence' which in themselves are semantically difficult to define (Abdul-Rahman and Hailes, 2000).

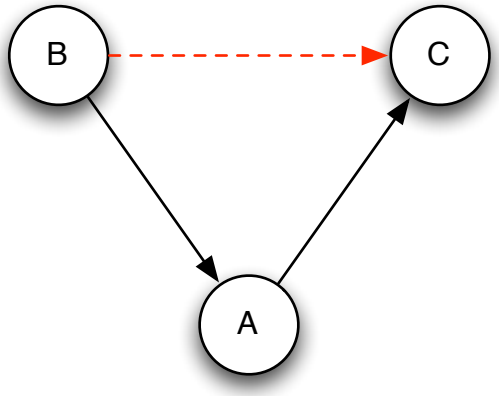


Figure 14: Inferred trust: B has no experience of C and so the trust is inferred through A.

2.3.3.2 Web of Trust

Several of the trust models that are discussed by Sabater and Sierra (2005) use a propagated system of reputation where agents build up a level of trust in other agents by querying their established contacts (Abdul-Rahman and Hailes, 2000; Carter et al., 2002; Schillo et al., 2000). The idea here is to help cope with the *sparsity problem* which occurs in large online systems; with a very large network of users, most of whom are engaged in one-time interactions, it is hard for any one agent to build up a trust rating for every user. Instead they must rely on trust propagated through others, as illustrated in Figure 14. This system is known as a *'web of trust'*. Abdul-Rahman (1997) implemented the first specific computational version, but it was originally proposed by Zimmermann (1994) in the context of the PGP security protocol:

> *"As time goes on, you will accumulate keys from other people that you may want to designate as trusted introducers. Everyone else will each choose their*

own trusted introducers. And everyone will gradually accumulate and distribute with their key a collection of certifying signatures from other people, with the expectation that anyone receiving it will trust at least one or two of the signatures. This will cause the emergence of a decentralized fault-tolerant web of confidence for all public keys."

A web of trust approach has been used in many different approaches to assigning trust across a network of agents. O'Donovan and Smyth (2005) used it in implementing an improved recommendation algorithm for movies, incorporating trust so as to augment the traditional rating-similarity approach. Trustmail (Golbeck et al., 2003) aims to improve the filtering and flagging of both important mail and spam through examining the network of acquaintances of both sender and receiver. Windley et al. (2007) look to provide an automatic moderator system for blogs based on the reputations of the people leaving comments.

In the context of our own work, we aim to implement a weighting scheme for authors based on the co-occurrence of annotations within both threads and web pages. This view of a web of trust is combined with ranking algorithms which may be seen as akin to the trust values themselves. PageRank (Page et al., 1998) provides a confidence value for each page based on the pages which link to it, the idea being that there is an implicit 'trust' that the author of a good page would only link to a good page. This idea was developed further by Gyöngyi et al. (2004) who implemented TrustRank.

While many of the algorithms discussed incorporate some measure of cheating or falsification, Guha et al. (2004) are one of the only authors to actively attempt the propagation of distrust. This can introduce many problems such as the interpretation of negative probabilities and zero values. We choose to ignore the idea of distrust and instead concentrate on trust. The trust that we shall be modeling however, is not contingent on global values but remains local. Incorporated into the algorithms we have created (see Chapter 4) is an idea of trust akin to Carter et al. (2002); contributions to society are important. We view this contribution from the standpoint of how much an agent can engender conversation between other agents within the society. In the next section we discuss the measurement and assessment of the quality of that conversation.

2.4 Data Quality

Data Quality is concerned with the quality of data which is collected, stored and used. As we shall see in Section 2.4.3, there are many factors which need to be taken into account when deciding upon the quality of data. There are also many different opinions on which factors should be taken into account and which should be ignored. Before looking at the quality of data however, it is necessary to attempt to define the term

'data' itself. We say attempt, as we wish to reflect the surprising difficulty which arises in creating a definition for data.

2.4.1 Defining Data

Mealy commented that *"We do not, it seems, have a very clear and commonly agreed upon set of notions about data – either what they are, how they should be cared for, or their relation to the design of programming languages and operating systems."* (Mealy, 1967). Many different definitions and classifications have been presented in the intervening years, some more comprehensive than others. In an attempt to skirt the issue, several authors have opted to use the terms data and information interchangeably. This is very confusing however; some use the simple classification of data as *'Any kind of information which is analysed systematically'* (Dasu and Johnson, 2003). In contrast, information is then seen as *"processed data"* (Wang et al., 2001) or *"any kind of knowledge or message that can be used to make possible a decision or action"* (Langefors, 1973). From these definitions alone we can see the circular logic and contradiction which has caused so much confusion in the past.

Redman (1997) provides an extensive overview of the competing ideas of what data is and how it should best be defined. His requirements state that data be defined in a way which is clear and simple, has no mention of information (so as to avoid circular logic), agrees with common usage, is comprehensive (embracing both representational and conceptual facets), is widely applicable and intuitively suggests quality dimensions of the data.

2.4.1.1 Defining Information

Again, Redman (1997) provides an excellent discussion on what constitutes information. Redman views data as signals, pointing out that this gives wider scope than 'messages' since messages implies an active role in the creation of the data. Signals may be sent out from inanimate objects too, and it is the role of the observer to process these signals. Since it is not possible for an observer to single out any one signal (bearing in mind that the very act of interpretation involves the recollection of past experiences and therefore signals), we must consider a collection of signals. Information is then defined as the non-redundant part of this collection of signals, which by definition is informative and therefore 'information'. Redman does point out the inherent uncertainty in this definition, it being reliant on the observer and their past experiences.

2.4.1.2 Types of Data

Although we may be able to define data, there are still many types of data which sit under this definition. Data may be rigidly structured, as is the case with relational databases and the data they house. It may also be semi-structured; the most prevalent sort of data of this type is XML data which may or may not have a schema associated with it. As a result, the same information may be represented in several different ways (e.g addresses which may be given to varying degrees of detail, or represented as a single field, or several distinct sub-fields). Lastly data may also be completely unstructured as, say, the transcript of a conversation or free-text.

Data may also come from many different sources. When drawing parallels between data manufacturing and product manufacturing, data may be seen as a raw material, a component material (stored for a short period and discarded once an information product is created), or as an information product itself. This last classification is troubling in that it again creates an ambiguity about *data* vs. *information* (Wang et al., 1998).

Most relevant to this thesis is federated data. This is data that comes from several sources and can require disparate data to be combined in an approximated manner. Web data is federated, especially user-generated content which not only comes from different web pages, but also from many different authors. There is also a lack of control on what format this data will be in, varying from structured data (tags as discussed in Section 3.1.2 may be thought of as structured since they are inherently of a single form, being a single word) to unstructured (free-form annotations). In this thesis we shall be looking at semi-structured data in the form of XML annotations.

We should also take into account the changing nature of data with respect to time. Temporally, data may change from one form to another. It is important to make the following distinctions when considering data quality as the quality of data may change with the data itself; data may be thought of as stable if it is unchanging and constant with respect to time, an example being a person's date of birth, or publication dates; data may be long-term changing or frequently changing, however this distinction is domain dependant.

2.4.2 Data Systems

Since the way in which data is collected, stored, represented and used can be quite repetitive, Redman refers to the life-cycle of data. The way in which this cycle progresses and changes the state of data can be modeled in a system, the focus of which helps to define the system type. Figure 15 shows the data cycle for two distinct types of system. A distinction is necessary as there are many situations where acquisition and usage are

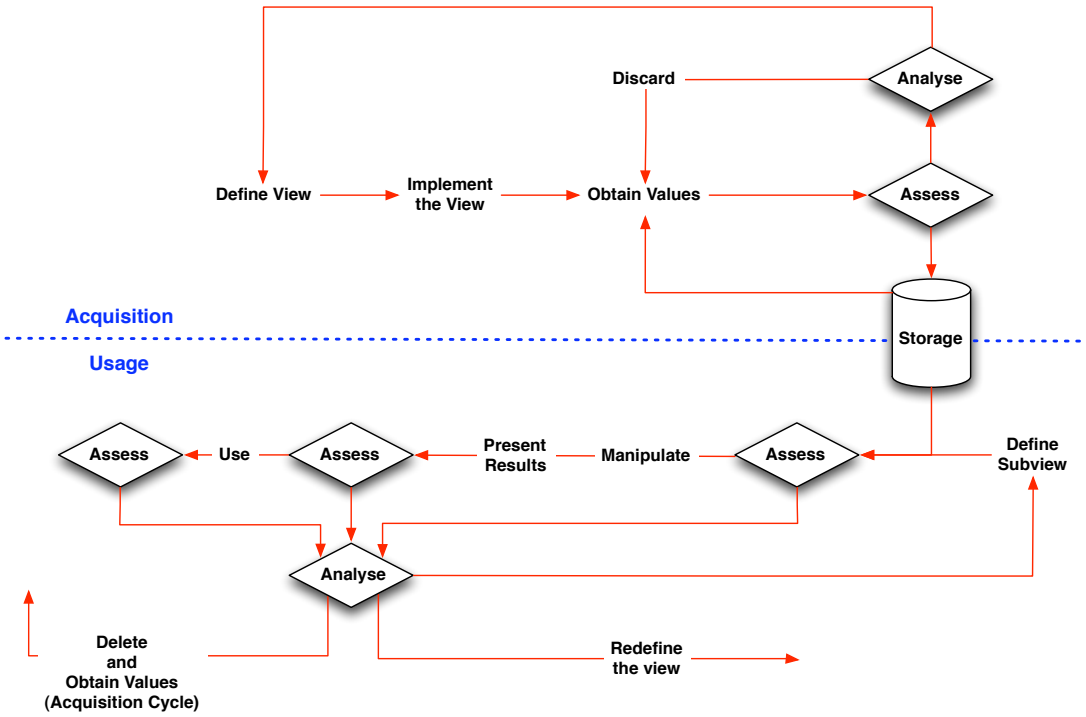


Figure 15: The data Systems related to Acquisition and Usage Cycles.

performed by different systems (e.g. in market research when data collection is handled by a separate and specialised company different to the company which has commissioned the research).

If a system is mostly concerned with the acquisition and storage of data, then it is said to be of an *acquisition type*. The main stages of the data acquisition cycle may be seen in the top half of Figure 15. Before acquiring any data, an appropriate view of the system must be decided upon; what is the aim of the system and what data must be captured? As we shall see in Section 2.4.3, the elimination of redundant or contradictory data is key to data quality. Implementation is a case of schema definition, representational consistency, and taking into account requirements and limitations of the storage method. Obtaining specific values is one area where many of the data quality issues seen today may be reduced, making it a vital part of any data acquisition cycle. Obtaining incorrect values here can result in misinterpretation of data due to poor data collection techniques. Updating of records is finally achieved through the addition of new data, removal of old data and modification of existing data.

Following on from the definitions of data and information provided in Section 2.4.1 we can see that while the main goal of the acquisition cycle is the manipulation and transportation of data, the usage cycle deals more with information. To do so, an appropriate sub-view must be defined from which requirements for data usage are taken. This sub-view aims to utilise just a subset of the available data, much as the view defined

in the acquisition cycle aims to represent only a subset of the real-world from which the data was taken. Retrieval of data from storage is taken next and is closely coupled with the possible (and optional) manipulation of retrieved data. Lastly the data is presented to the data consumer/user who may or may not then use this data.

Through these two cycles of data acquisition and usage, much assessment is required. The aim is to assess the quality of the data being retrieved/used/stored etc. As we shall see in the next section, there are many different dimensions to data quality beyond the simple accuracy of the data. These assessment and analysis phases aim to discern whether or not these quality requirements have been met. If not, there may be issues with the way in which the data is being handled (resulting in the re-defining of the view/sub-view) or problems with data itself (such as its currency, detail or value).

2.4.3 Classifications of Data Quality

Data quality (DQ) has been defined in many ways having originally been considered to consist predominantly of the accuracy of the data being analysed. This over-simplification has been criticised for its lack of distinction between the different aspects of quality. Accuracy itself is also difficult to quantify as it is highly dependant on the domain of usage (Dey, 2001; Strong et al., 1997).

Redman (1997) divides the dimensions of data quality into 3 main categories; a conceptual view which is akin to the defining of a view or sub-view in Figure 15, requiring the definition of the subset of available dimensions in which to interpret the data; a view of data values and quality in relation to these values; a format or representational view of the data, dependant on storage-method limitations and schema requirements. Batini and Scannapieco (2006a) refer to this view as the "Intuitive approach" to data quality relying on common-sense and observation to define the dimensions of data quality. Table 1 provides the dimensions within each of the categories.

The conceptual dimensions of data quality cover many of the issues within the acquisition cycle of data. Defining an appropriate scope and level of detail after ensuring the relevance, obtainability and changeability of data are also acquisition issues. One must take into account the possibility of external factors influencing data quality as well as choosing a composition which is intuitive and minimal.

Two other major classification of data quality were presented by Wand and Wang (1996) and Wang and Strong (1996) respectively. The first of these is referred to as the "theoretic approach" to data quality. This approach considers an information system[8] as a representation of a real-world system. Quality is divided into just 5 aspects, each

[8] "An information system is modeled as a mapping from events in the world to signals. Users take actions based on the signals provided by the system."(Wand and Wang, 1996)

Table 1: Dimensions of Data Quality according to the Intuitive Approach. Quoted text taken from Redman (1997)

Category	*Dimension*	*Definition*
Conceptual	Content	"Relevance of data, obtainability[sic] of values, and clarity of definition"
	Scope	"The degree to which a view encompasses enough data to meet the needs of all applications and the amount of excess data"
	Level of Detail	"The level of data that must be included and how precise that data must be"
	Composition	"The internal structuring of the view ... characterised by naturalness, identifiability[sic], homogeneity and minimum redundancy"
	View Consistency	"Semantic and structural consistency"
	Reaction to Change	"The ability of the view to accommodate change"
Data Values	Accuracy	The nearness of a value v to some value v' in the attribute domain ... considered as correct
	Completeness	"Degree to which values are present in the data collection"
	Currency	"The degree to which a datum is up-to-date"
	Consistency	The same datum in overlapping collections is represented in a non-conflicting manner
Representational	Appropriateness	"One format is more appropriate than another if it is more suitable to users' needs"
	Interpretability	"User may easily interpret values correctly"
	Portability	"[The format] can be applied to as wide a range of situations as possible"
	Precision	"The ability to distinguish between elements in the domain that must be distinguished by users"
	Flexibility	Changes in user needs and recording methods may be easily accommodated
	Null Values	Able to represent null values and distinguish them from default and representable values
	Efficiency	Must use storage media efficiently without causing ambiguity or inconsistency
	Representational Consistency	Coherence and accordance of physical instances of data with their formats

specified by their negation:

- **Accuracy:** Inaccuracy implies the information system represents a real-world state different from the one that should have been represented.
- **Reliability:** Reliability indicates whether the data can be counted on to convey the right information.
- **Timeliness:** This refers only to the delay between a change of the real-world state and the resulting modification of the information system state. Lack of timeliness may lead to a state of the information system that reflects a past state of the real world.
- **Completeness:** Completeness is the ability of an information system to represent every meaningful state of the represented real world system.
- **Consistency:** If there is more than one state of the information system matching a state of the real system, inconsistency would mean that the representation mapping is one-to-many. Wand and Wang (1996) however, do not consider this a deficiency.

Wang and Strong (1996) provided a far more extensive categorisation breaking quality down into 4 main parts (Table 2): Intrinsic DQ, Accessibility DQ, Contextual DQ and Representational DQ. These 4 categories and 15 attributes have been whittled down from a starting point of 179 attributes which were complied by surveying 112 people. It had been pointed out that the intuitive approach allowed for the selection of "the most relevant attributes to a particular goal of study", and the theoretical approach allowed for the provision of "a comprehensive set of data quality attributes that are intrinsic to a data product", both failed to capture the data consumers'/users' needs.

2.4.3.1 Data Quality in the World Wide Web Domain

Parker et al. (2006) provide an excellent overview of the frameworks for data quality which has been proposed for this domain in the past. Of particular interest to us is that proposed by Zhu and Gauch (2000), which we feel is highly applicable to our needs. (For a more detailed explanation of how we incorporate the attributes proposed by Zhu and Gauch (2000), see page 123.) They give 6 attributes on which the quality of web pages (and implicitly the data within that web page) may be judged. We do not consider the *relevance* attribute:

- **Currency:** How recently a web page has been updated, measured as the time stamp of the last modification of the document.

Table 2: Dimensions of Data Quality according to the Empirical Approach (Batini and Scannapieco, 2006b)

Category	*Dimension*	*Definition (Extent to which . . .)*
Intrinsic DQ	Believability	"Data are accepted or regarded as true, real and credible"
	Accuracy	"Data are correct, reliable and certified free of error"
	Objectivity	"Data are unbiased and impartial"
	Reputation	"Data are trusted or highly regarded in terms of their source and content"
Contextual DQ	Value-Added	"Data are beneficial and provide advantages for their use"
	Relevancy	"Data are applicable and useful for the task at hand"
	Timeliness	"The age of the data is appropriate for the task at hand"
	Completeness	"Data are of sufficient depth, breadth, and scope for the task at hand of their source and content"
	Appropriate Amount of Data	"The quantity or volume of available data is appropriate of their source and content"
Representational DQ	Interpretability	"Data are in appropriate language and unit with clear data definitions"
	Ease of Understanding	"Data are clear without ambiguity and easily comprehended"
	Representational Consistency	"Data are always presented in the same format and are compatible with the previous data"
	Concise Representation	"Data are compactly represented without being overwhelmed"
Accessibility DQ	Accessibility	"Data are available or easily and quickly retrieved"
	Access Security	"Access to data can be restricted and hence kept secure"

- **Availability:** Calculated as the number of broken links on a web page divided by the total numbers of links it contains.
- **Information-to-Noise:** The proportion of useful information contained in a Web page of a given size meaning the ratio of the total length of the tokens after pre-processing divided by the original size of the document.
- **Authority:** The reputation of the organization that produced the Web page based on the Yahoo! Internet Life reviews[9].
- **Popularity:** How many other web pages point to this particular web page. Information on the number of in-links to a web page was taken from a 1999 snapshot of the AltaVista site[10].

2.4.3.2 Data Quality and User-Generated Content

Within the context of this thesis we have attempted to look at the quality of the data being provided by a federated web data source of users. The approach of Zhu and Gauch (2000) fits nicely with the ideas that we have for providing an automatic quality measure to the contributions/annotations of web users. Accessibility will be ignored as the issue of access does not arise in the scenarios which will be discussed. Within the sub-categories proposed by Zhu and Gauch (2000) however, there are some highly applicable ideas. We shall adapt the idea of quality measures for an entire web page to take into account instead the annotations provided to a web page.

2.5 Annotation

While reading is an inherently passive activity, writing requires far more effort on the part of the writer. It is perhaps not surprising then that "the most pervasive activity around documents is reading" (Brush et al., 2001). Brush et al. (2001) also note that the act of reading is in fact closely followed by annotating. Annotation forms a bridge between the separate activities of reading and writing, allowing the reader to take a more active role in the creation and dissemination of information. This *active reading* role (Adler, 1972) is something which has become more prominent with the advent of e-books, Web 2.0 and specific digital annotation software.

Definition 1 ***Annotation***[11]

i) A critical or explanatory commentary or analysis

[9] http://www.zdnet.com/yil
[10] http://www.altavista.com
[11] *http://en.wiktionary.org/wiki/annotation*

ii) A comment added to a text

iii) The process of writing such comment or commentary

iv) (computing) Meta-data added to a document or program

v) (genetics) Information relating to the genetic structure of sequences of bases

From Latin annotātiōnem, accusative singular of annotātiō ("remark, annotation"), from annotātus, perfect passive participle of annotō ("note down, remark").

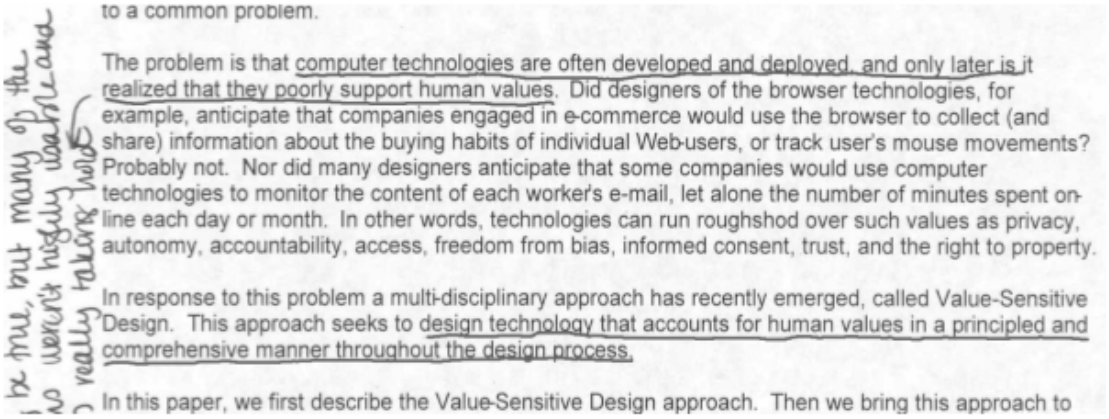
Annotation and the act of annotating manifests itself in many different manners and for many different reasons. The annotation taxonomy presented by Ovsiannikov et al. (1999) states that annotations may be created 'to remember, to think, to clarify, to share'. An annotation may be of any modality, be that audio (as in a sound-byte or song), visual (photographs or video) or most commonly written. It is usually the case that the annotation itself is however in the same modality as the document or source which is being annotated (Agosti et al., 2007). As is the definition, the purpose of an annotation is to provide additional explanation or clarification to the annotated source. In doing so, a symbiotic relationship is created between annotation and annotated object with the information in each re-enforcing and benefiting the other.

While the purpose of an annotation may be to clarify and provide information to an annotated source, the method of annotating can vary greatly. Annotations can be highly transient in nature, marking out a reader's current state-of-mind when reading a document. On the other hand, the persistent and permanent nature of an annotation can lead to its usefulness growing. An annotation may aid in data-provenance helping to preserve information on the origins of a document, as well as interpretations of semantics, and adding contextual information. Data provenance is *"the description of the origins of data and the process by which it arrived [in the database]"* (Buneman et al., 2001). Marshall (1998) provides a thorough overview of many of the different way in which annotation may be used.

2.5.1 Physical Vs. Digital

As mentioned annotations may be created in any modality, but are mostly frequently found in the same modality as the annotated source. Most research has focussed on written documents annotated by written annotations. In the physical world these annotations take the form of underlining, margilinia, highlighting etc. with the exact method specific to each annotator. The vast majority of these annotations are anchored to specific points (phrases, words, paragraphs) within the source documents (see Figure 16). This is mainly due to the increased effort required on the part of the annotator to

recreate the context of an annotation which is recorded separately to its corresponding document (Brush et al., 2002; Marshall, 1997). This increased effort in fact leads to a different style of annotation in which the information contained within the annotation is of a more general nature, recapping or summarising the document in full. It has also been noted that the use of anchored annotations leads to increased creation of new information about the annotated document, as opposed to summaries of pre-existing information (Wolfe, 2000).


to a common problem.

The problem is that computer technologies are often developed and deployed, and only later is it realized that they poorly support human values. Did designers of the browser technologies, for example, anticipate that companies engaged in e-commerce would use the browser to collect (and share) information about the buying habits of individual Web-users, or track user's mouse movements? Probably not. Nor did many designers anticipate that some companies would use computer technologies to monitor the content of each worker's e-mail, let alone the number of minutes spent on-line each day or month. In other words, technologies can run roughshod over such values as privacy, autonomy, accountability, access, freedom from bias, informed consent, trust, and the right to property.

In response to this problem a multi-disciplinary approach has recently emerged, called Value-Sensitive Design. This approach seeks to design technology that accounts for human values in a principled and comprehensive manner throughout the design process.

In this paper, we first describe the Value-Sensitive Design approach. Then we bring this approach to

Figure 16: Annotations made on a text-book by a student (Marshall and Brush, 2004).

Annotations created in the physical world are often done so in a private manner and are not appropriate for public use. The style of annotation may not lend itself to being easily understandable by anyone but the original author. This is the case for annotations such *"No!"* or *"Must think on this"*. These styles of annotation are not self-explanatory and can also suffer from 'crises of intelligibility' (Marshall and Brush, 2004). When annotations move from a private to a public nature, they can lose their meaning since they are explicitly personal in nature and not designed to be of any use to persons other than the original author. This is not always the case however, and it has been shown that in some cases private annotation can indeed be of use to the public (Marshall, 1997; Shipman et al., 2003).

Digital annotations mirror the annotations of the physical world, allowing people to become more involved in the authoring process. One large advantage of digital annotations however is that they may be organised and searched if desired; this allows for the creation of an 'annotation index' (Ovsiannikov et al., 1999) from which information may be retrieved. This index allows for the serendipitous discovery of annotations and annotated documents which may relate to the current context. This index (and annotations in general) provides the necessary information to create a summary of important information which is far less author-centric in nature. By observing where annotations are taking place, we can discover the information which is of most use to past (and by inference future readers of a document).

Within the digital document paradigm, public and private annotations are created through the use of access rights. There is also the possibility to share annotation with only those users with suitable access rights. This has a disadvantage in that the access rights are not going to decay with time, unlike in the physical case where a private document may be traded or sold and so all annotations contained within become shared/public. This does however reduce the opportunity to create experiments similar to those of Marshall (1997).

2.5.2 Annotations as Queries

Annotations focus the attention of future readers, allowing them to see what previous readers found of interest and importance. As well as this annotations may be leveraged to provide additional benefit to the current reader. Using annotations and annotated text as a means to query a collection of documents (or the web as a whole), we may find other documents which are of interest to the reader in their current context (Schilit et al., 1998). Annotations have been shown to provide better results than automatically selected text for relevance feedback (Golovchinsky, 1997), annotations being of smaller size than the entire document which is usually taken as context for traditional relevance feedback (Salton and McGill, 1986). Annotations more accurately reflect the intentions of the reader as opposed to traditional relevance feedback approaches, which while being statistically appropriate may not fully capture a user's intent in annotating (Golovchinsky et al., 1999). Annotations may be explicitly sent as queries to a search engine or in a query-less manner as described above, creating hypertext links between documents and enabling the reader to move between papers due to the annotations they have made.

2.5.3 Annotations as Hyper-links

While the query-less use of annotations above provides a means of linking documents, the use of annotations as hyper-links themselves is not covered. The explicit creation of hyper-links through annotations means that users can deliberately connect different documents. The difference from the approach mentioned in the previous section is that these links are not created automatically, but instead are made manually by the annotator themselves. These links can help to further clarify the information within an in-context annotation, or may be essential in connecting an annotation which is stored separately to the source document. Again the use of annotations helps convey what readers of the source document believe to be important rather than simply what the writer regards as important. It is also possible for information which was not available at the time of writing to be added in this manner.

As mentioned in Section 2.5.1, the use of anchored annotations such as hypertext

links causes a more discursive style of annotation. Annotations can be threaded, attaching annotations to each other in the form of a conversation. In this way annotations about annotations may be created and held in temporal order (Lanagan and Smeaton, 2007, 2009). As noted by Zheng et al. (2006) *"The full power of structured annotation lies in the interplay between normal workflow (editing, commenting, and reviewing) and the ability to capture that workflow and use it to manage future workflow"*.

2.5.4 Grouping Annotations

It may also be desirable to group annotations together into bundles or clumps of related annotations (Ovsiannikov et al., 1999; Zheng et al., 2006). This approach helps to alleviate the difficulty of searching annotations due to annotation length which is generally quite short. Grouped annotations may be thought of as similar in some way, and can be combined to form a pseudo-document from which relevant information may be retrieved (Abel et al., 2007). Groups may also be created using automatic filtering techniques, allowing for temporal, length, user-specific etc. filtering of annotations. In this way a person can review all annotations which have happened since last viewing a document, or even just the annotations of a particular person. This grouping and filtering, while possible on physical annotations, is far easier with digital annotations. Within the physical domain filtering relies on such visual queues as handwriting, color-coding and style of annotation (underlining etc.) to differentiate and group annotations (Marshall, 1997, 1998).

The worth of annotations as information in their own right has been discussed (Agosti and Ferro, 2003), the annotations being autonomous from the document they annotate but retaining some sort of link. While annotations help to enrich a document providing a focus to readers, they may also serve as a springboard to new ideas and documents. By collecting a document's annotations together (and essentially creating a new document in the process), one may be able to construct a new document based on the annotations. In this way the annotations have retained a link to the original source document, whilst becoming a document in their own right (Bottoni et al., 2003).

The amount of novel information provided by annotations may also affect their representation. One can easily imagine a situation where annotations of particular interest to a user may be presented in a different manner. The degree of *semantic distance* (Smeaton and Quigley, 1996) between an annotation and source may range from 0 (being a highlighting or underlining of text) to 1(representing a completely unrelated jotting or note created by a user e.g. 'Time for lunch'). It is difficult to equate this semantic distance exactly with worth of an annotation, but it does give some gauge as to how different the two sources are.

Location is also an important consideration with annotations, and may be used to aid in search tasks (Frommholz et al., 2003). In this thesis the position of annotations within threads is taken as a guide to annotation relevance. The position of annotations on a physical page is also discussed in (Marshall, 1997) when trying to digitise the annotations of users. This digitisation also raises the question of the usefulness of person/private annotations in a public context as mentioned in Section 2.5.1.

We have presented the background to the work in this thesis. In the next chapter we shall look at two of the systems that we have created which enable the creation of an annotations corpus, achieved through the adoption of Web 2.0 technologies.

CHAPTER III

WEB 2.0

Users of the web are becoming less content with the current publishing model employed on the internet (Hermida and Thurman, 2008). This model is taken from traditional publishing paradigms which do not afford the opportunity of readership interaction and participation. It is a "push" model where the author or creator of information pushes out information to an audience confined to reading. Readers/consumers play no active role in this process, but instead remain passive in both the creation and dissemination of information. Some publications are attempting to move away from this model and allow for more interaction[1].

In this chapter we consider many of the new methods of interaction amongst internet users collectively called *Web 2.0*. We give an overview of the state-of-the-art in Web 2.0 research and application. Following this, we outline two systems which we have developed to study and explore the state-of-the-art.

3.1 Web 2.0: People Talking to People

Though sometimes derided as "marketing hype" or "buzz words" due to the lack of an exact definition, Web 2.0 may be seen as an attempt to address some of the limitations of the original web. Conceived in a brainstorming session between O'Reilly and MediaLive International, the idea of Web 2.0 grew out of the remnants of the dot-com bubble of 2001 (O'Reilly, 2005). Noting that several companies had managed to prosper while all around were collapsing, the session panel believed that these companies and web-sites

[1] http://www.usatoday.com/news/2007-03-02-editors-note_N.htm

Figure 17: A tag cloud of the most common Web 2.0 terminology

must have had something in common.

One of the main ideas behind Web 2.0 was changing the push model of publishing to one where the audience of information consumers might be able to contribute. Much of the technology used and re-branded as Web 2.0 had in fact existed for a long time, another reason for sceptics to rally against the re-branding. The growth of internet and online community however, meant that possibilities for user participation and interaction had increased considerably. In the following sections, we highlight some of the main uses of Web 2.0, giving examples of current commercial implementations as well as research which has focussed on the same area. A more thorough discussion of the ideas presented here may be found in Chapter 2.

3.1.1 Vote for me

Collaborative filtering involves the mining of past user choices to improve the experience of other users. Tapestry was the first system to employ the idea of collaborative filtering, helping users to filter a growing number of e-mails for the most interesting and useful ones (Goldberg et al., 1992). Users were able to annotate mails, providing other users with a means of filtering the messages which they received in future. The system relied

on there being two types of users; the eager annotators who would read the majority of messages, providing annotations to each; and the users who would wait for these annotators to provide a guide as to what was useful. While no specific voting scheme is in operation, by combining a large number of filters based on several users it would technically be possible to create a form of voting. The more users marking a message useful/interesting, the higher the message is "voted".

A more obvious form of voting is employed by the GroupLens project (Konstan et al., 1997), where the recommendations of other users are aggregated so as to provide a score for Newsnet postings. This work extended the work of Tapestry by removing the requirement of the user to choose which filters he/she wished to use. The GroupLens system provided a means of combining the ratings given to posts, meaning that the identity of the rater was of little consequence. Users were able to rate Newsnet posts anonymously and still remained of value to the system. A disadvantage of anonymity is the loss of user information which may be leveraged to create a more tailored experience.

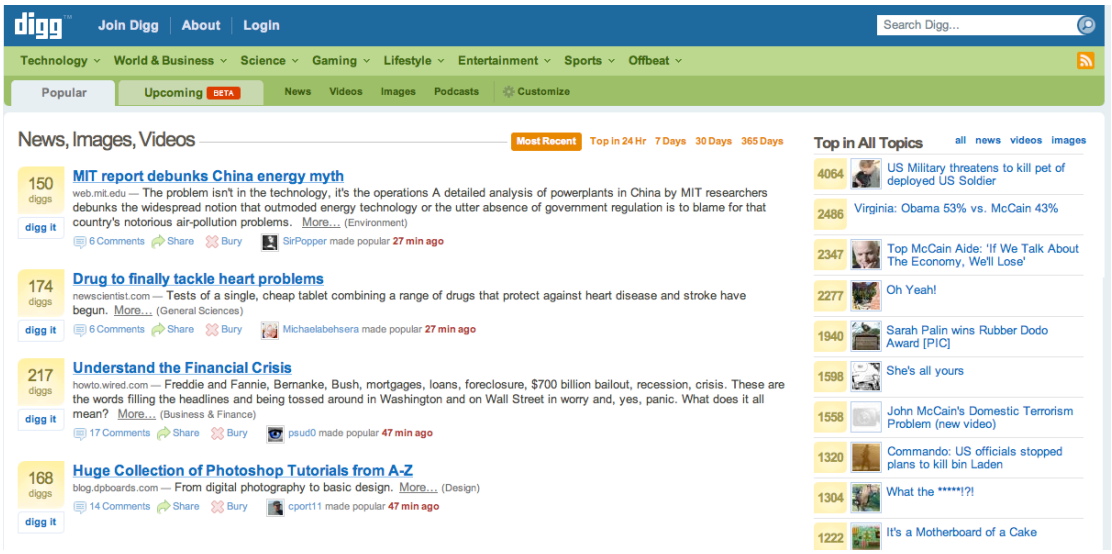


Figure 18: The front page of Digg showing those stories which have been voted for the most.

One of the most popular Web 2.0 social rating sites is Digg[2]. The site allows users to post a link to web-pages, podcasts and digital content which they find interesting for users to vote on (i.e. the post is 'dugg' by other users). In order to vote for a posted link, users must be registered with the site. The more people who vote for a post, the higher it is placed on the site's ranked list of posted 'diggs' with the goal being to have the post appear on the front/first page of the web-site as shown in Figure 18. Within the site, users are able to create lists of other users who they wish to follow, being notified any time a post is made by those users whom they are following. This means that acquiring a lot of followers can lead to increased influence within the site. Lerman

[2]http://www.digg.com/tour/

(2007) has shown that friends, or followers of other users, prefer to dig the posts of each other which can lead to "tyranny of the minority". Unlike GroupLens, users' identities are explicit and so while ratings are effectively anonymous, the benefit of posts being 'dugg' is past on to the user who first posted the item; the more often a post appears on the front page, the greater the likelihood of increasing the number of followers and prestige.

Within the context of the two systems that we shall present in Section 3.2, the prestige or value of users to the community as a whole is measured not by voting but in terms of interaction. This idea has been used in the past to provide users with a trust score or reputation (Zacharia et al., 2000; Windley et al., 2007), enabling prioritisation of users and limitation of privileges when using web-sites. The idea of trust and reputation is widely used in sites such as eBay[3] to provide users who have never interacted with knowledge based on the past interactions of each user with the community. Terveen et. al (Terveen et al., 1997) note that *"the distinct number of recommenders of a source is a plausible measure of resource quality"*. Within our two systems we make the assumption that interaction is a form of recommendation for a user. By this we mean that a user who is able to create conversation between a large number of distinct users has provided something of value to the community as a whole. The increased number of participants is seen as a measure not of direct quality of the conversation being had, but of the level of interest created. The more interest, the more value.

3.1.2 Tag, you're it.

Tagging of content refers to the application of single words (or concatenation of words e.g. "stateOfTheArt") to objects on the web to make identification easier. Originally used by the photo-sharing web-site Flickr[4] to alleviate the problem of searching for photos (Figure 19), tagging has become a highly active area of research. Social bookmarking site del.icio.us[5] have seen rapid growth in users since its introduction in 2003 and now has over 5 million users and 150 million tags. These massive data-sets has been the focus of research into "folksonomies" or social tagging (Nov et al., 2008; Paolillo and Penumarthy, 2007).

Folksonomy is a portmanteau of the words folk and taxonomy, created by a collection of users. It has been noted that organised ontologies may arise from the seemingly chaotic assignment of tags to resources by an uncontrolled and unrestricted user community (Mika, 2007). Folksonomies have been shown to aid in the retrieval process, both by providing keywords with which to search, and by using the tags given to a resource

[3] http://www.ebay.com
[4] http://flickr.com
[5] http://del.icio.us

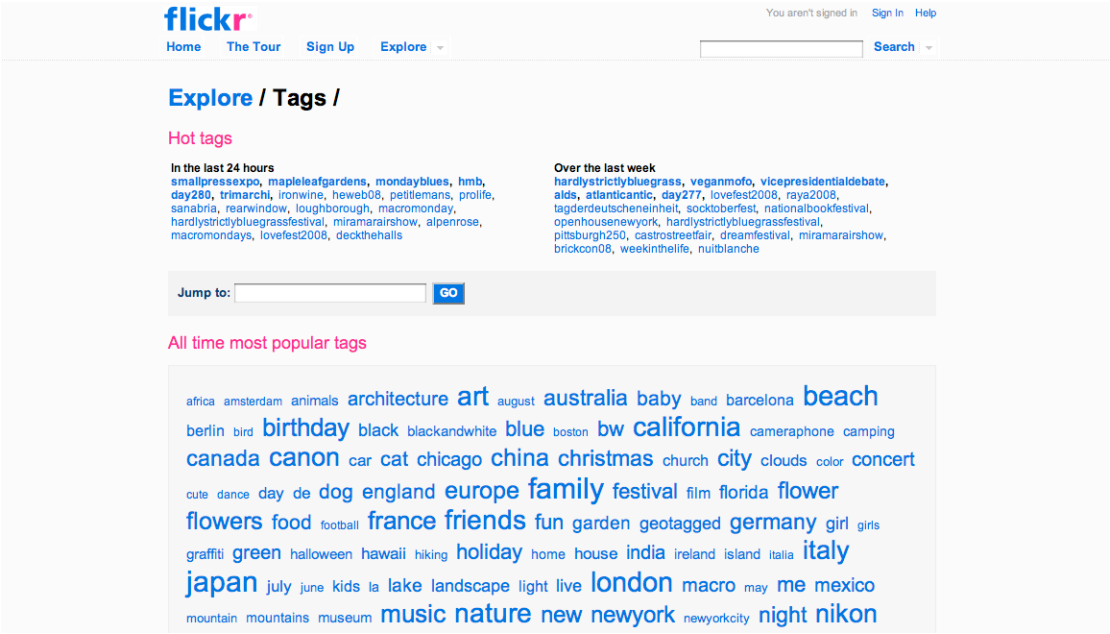


Figure 19: Lists of tags which have been applied to photos in Flickr to aid in browsing and retrieval.

to provide context for a query (Hotho et al., 2006). Folksonomies are also becoming one of the focus-points of efforts to create the Semantic Web (Berners-Lee et al., 2001) due to the fact that the tags are created by users. As such, these tags provide a far better description of the resources which they describe than anything created by a machine (Wu et al., 2006; Mika, 2007).

Folksonomies and tags have been used to aid in the browsing process also, guiding users to content which is more relevant to their current information needs. Tags have been shown to aid users in navigating between sites which, while not necessarily hyper-linked, are semantically relevant to each other. The additional information provided by tags can help to organise results returned by an initial search, providing a more focussed and coherent browsing experience (Li et al., 2007; Millen and Feinberg, 2006). The value of tags is even more evident when browsing visual media. In conjunction with visual features such as texture and colour, tags can provide the additional information required to provide meaningful results to image queries (Aurnhammer et al., 2006). An example of this is a search for "beaches" which when performed against tags seems easy, visually however a vast array of problems are encountered. One can imagine any picture showing predominantly yellow colours at the bottom and blues at the top would be returned.

In the two systems that we have created, we have not provided the ability to create tags. The main reason for this was that we believe there is no real requirement for such tags, since the material provided is of a focussed and consistent nature. As stated, the benefit of tagging can be seen in the additional information and organisation which tags bring to a diverse and expansive collection of media. In our case the media itself is

focussed on a single sport (and single sporting event), and so many of the most common tags found on flickr (in the case of SportsAnno these include "worldcup", "Germany", "football") would apply to every piece of media content in its respective corpus. While some tags may have been applied to differentiate between, say, teams we felt that the size of the corpus made this unnecessary. A version of tags was applied to images within the Annoby experiments which will be discussed in Section 3.2.3.

3.1.3 Social Commentary

A major drawback of tagging is the fact that tags are single words. They lack depth of expression or explanation. The concatenation of words to form single tags (e.g. 'high-school' or 'creditcrunch') goes some way to alleviating this problem. Unless additional indicators such as the co-occurrence of tags is taken into account however, there is no real way to differentiate two piece of media tagged with the same or similar tags. The reason for tagging is lost, although it may in some cases be obvious (as is the case with noun tags) (Golder and Huberman, 2006; Begelman et al., 2006).

Free-text annotations or comments can be used to give a more descriptive and semantically accurate impression of information which is being annotated. Though not as prevalent as tagging, annotation systems such as socialbrowse[6] and i-Markup[7] allow for free annotation of part or all of a web-page. The first of these two systems is in fact aimed at allowing users to create threads of conversation in realtime, making the browsing experience a far more social activity. It has been shown that in the same way they prefer to 'dig' their friends, users will visit web-sites that others have visited in the past. ASSIST (Freyne et al., 2007) built on these assumptions, enabling users to see where others had browsed before them, though not allowing for any actual annotation. OATS (Bateman et al., 2006) was designed to allow students to create and share annotations on course-work, augmenting the idea of tagging with free-text annotations. While the tags provide a means of categorising and clustering annotations, the annotations themselves provide the information.

ASSIST and OATS both allow for the annotation of documents and parts of these documents with free-text annotations. This style of annotation is quite recent, and has become more popular with the advent of Web 2.0 and social networking. Earlier forms of annotation or commenting were restricted to web-forums and web-logs (blogs). This form of commenting has begun to be seen on main-stream web-sites such as BBC news[8] and YouTube[9], enabling users to comment on existing media. It is interesting that newspaper web-sites and dedicated internet news web-sites are beginning to allow this

[6] http://socialbrowse.com/
[7] http://imarkup.com/download/plugin/server_plugin.asp
[8] http://news.bbc.co.uk/
[9] http://www.youtube.com

form of annotation as it is exactly these publishing bodies who had created the "push" form of media.

Annotations also form the basis of the two systems we have developed, allowing users to conduct threaded in-context discussions about sports video and associated newspaper reports. 'In-context' refers to the anchoring of comments in place and neighboring the phrases/quotation to which they refer within the original document instead of being placed in a separate area. O'Hara and Sellen (1997) notes that the smooth integration of annotation and reading is an essential and vital quality of any annotation system, something which in-context annotation provides. As shall be shown, this additional information present in the annotations helps to focus users on the information which is of most interest to the user community.

3.2 Two novel Web 2.0 systems

In order to study the usefulness and potential of user-generated content to aid in the information retrieval and browsing process we have created two annotation systems. Each of these systems focussed on a globally recognised and high-profile sporting event, providing the systems' users with a means of viewing and discussing the related broadcast media. Additional to this, all discussion was recorded as permanent annotations to the media and presented in-context to subsequent users of the system. These systems aimed to build on present technologies, bringing together disparate strands of the viewing and sharing process.

With the proliferation of sports video and media on the Internet, sports channels can now offer live web casts of matches as well as recorded footage. Along with this video comes large numbers of reports written to capture all the major events within a match. Since these written reports are essentially designed as a summary of the matches they describe, they may be used as a guide or key into the recorded video. As it stands these reports and streaming footage are very disjointed with no possibility for a user to simultaneously browse both written match reports and the associated video media. Beyond this if a user wishes to comment on events within a game, he/she must go to a third resource, a forum say, to be able to actively post an opinion or point of view. This loss of context and need to reference the original material requires a great deal of effort on the part of the user. More sites are beginning to see this problem and address this by allowing users to post comments at the bottom of articles published on the site. Comments posted in this fashion however tend to be of a general nature, recapping the documents to which they are attached (Brush et al., 2002). They are less discursive in nature.

The advantage of the systems we have developed lies in their ability to bring together

and combine 3 currently separate aspects of sports recording. Users are able to read match reports taken from newspaper web-sites, view the match video associated with the reports and create in-context comments which are then used as the basis for discussion amongst users of the system.

The immediate and easy access to both visual and written media, coupled with the ability to leave comments within the text for other users leads to a more directed and communal style of annotation. Since video is always present, it is possible for a user, at any time, to see the arguments presented in writing first-hand and to couple it directly with the video. There is also the means to provide direct input into any discussions. The novelty of these systems is the opportunities they offer to become up-to-date with any talking points and also to contribute easily to any on-going discussion.

3.2.1 SportsAnno

The SportsAnno system was designed to give its users a comprehensive summary of all the action from the FIFA World Cup 2006 held in Germany. The aims of the system were:

- To allow users to become knowledgeable and form opinions about a sports event which they may not have seen live and be able to back up these opinions with visible evidence.
- To promote discussion about the sporting events and allow for the introduction of additional knowledge through this discussion.

Throughout the summer of 2006, all televised games were recorded and automatically marked up using event detection algorithms. At the same time, several newspaper web-sites were automatically scraped to obtain the corresponding reports for each of the games. The aim of the system is to give users the opportunity to voice their own opinions about all the events in the competition, with all the evidence before them. The FIFA World Cup was chosen for its huge appeal and as stated above, since sports can be a very polarising, it was thought that this type of material would produce the most discussion. Another large advantage of using the FIFA World Cup is the enormous number of written reports that accompany each match, leading to many different viewpoints even within the official media. The reports were chosen to give a cross-section of this opinion.

All sports reports in the media are in theory objective in nature, but this is never truly the case. Every report has an angle and the author, through their use of language, always portrays a certain bias (Tannenbaum and Noah, 1959; Wann et al., 1997). Sport has always been a highly contentious topic with each person having his/her own opinions

about the events that take place. SportsAnno is designed to capture such conversation and present it to its community of users. This is done in such a way as to promote further conversation.

SportsAnno is a closed system requiring users to first register before being allowed access it. When first accessing the system, users are presented with a list of all available games. Together with this, each game shows the number of comments already made, the number of new comments since the user last logged-in, and a short description of the match as shown in Figure 20. This description was chosen to be the subtitle from the BBC report web-page and proved to be an adequate guide to the match.

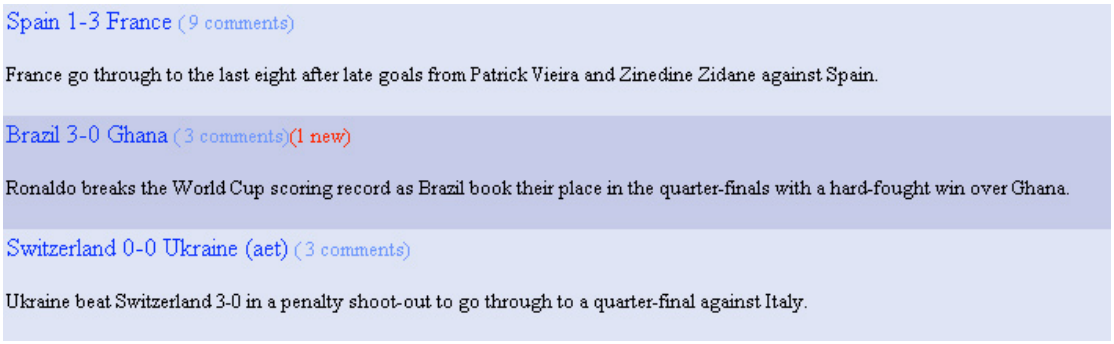


Figure 20: Users are made aware of where the activity has been since last logging in. This helps to focus attention on the areas of greatest community interest.

Once a game has been chosen, the user is presented with the full browsing interface allowing him/her to browse the reports and comments left by other users. This is shown in Figure 21. On the left of the screen the list of games is again available for easy navigation between matches. There are two major points of focus within the interface corresponding to the two complementary information sources: a collection of keyframes (representing the major events within the video) and the reports panel.

To the right are the keyframes, representing each of the segments within the video that have been marked as containing interesting events. Clicking on any keyframe will start playback from the beginning of the relevant segment. Each keyframe also has a small caption showing the time at which the segment starts. This is done so as to provide readers with an obvious correlation between the events within the written report and the events in the video itself. Discrepancies arise between the time point within the video and the actual time displayed on the in-game clock. This is most often due to extra time played out at the end of each half or injuries during play. For this reason, a tilde is added to times after 45 minutes (half-time) to indicate approximation. Since times stated within the reports are never to the nearest second but rather at a minute level, this slight inaccuracy was seen as no great inconvenience to the user.

The most important element of the interface is the reports pane (Figure 22). Placed centrally, it is here that users both read and annotate the newspaper reports. It is a

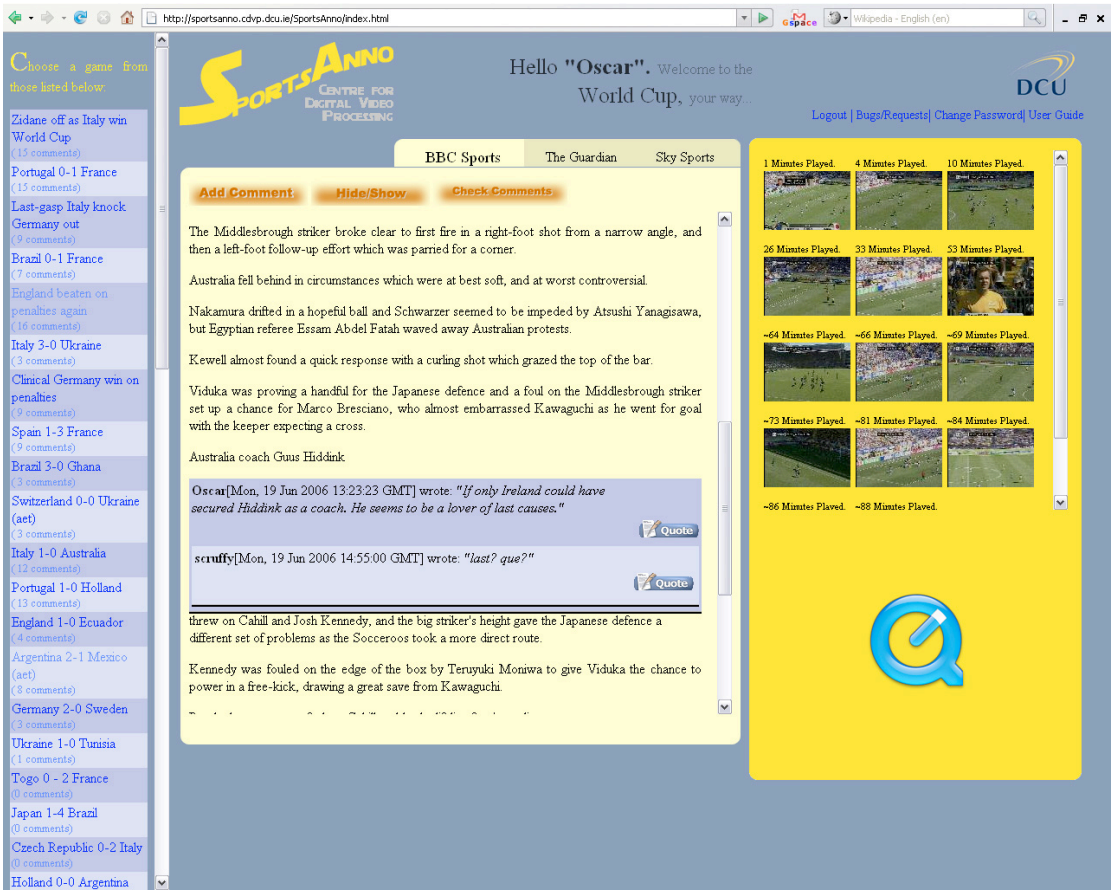


Figure 21: The SportsAnno main interface

tabbed panel collecting all the reports into one place. The reports are shown with all comments made by users placed in-context within the reports. It is possible to hide these comments by clicking the button at the top of the reports pane so as to read the report more easily. By default however, all comments are shown.

It has been found that the loss of context whilst annotating can cause the focus of conversation to change (Brush et al., 2002). Indeed, in-context commentary can allow for comments of a more specific and directed nature, as opposed to more general commenting on documents or events as a whole. It was for this reason that we chose to feature in-context comments within SportsAnno.

Commenting and the threads of conversation these comments promote, are the focus of SportsAnno. It was therefore of great importance to make the commenting facility as easy and intuitive as possible. In order to place a comment within a report, a user has simply to highlight a phrase within the report and click on the "Add Comment" button at the top of the reports pane. Commenting was restricted to phrases within a single paragraph (or a whole paragraph) so as to encourage discussion of specific points within the report. To reply to any comment posts, a user may click on the "Quote" button at the bottom of each post. This creates a thread anchored to the comment.

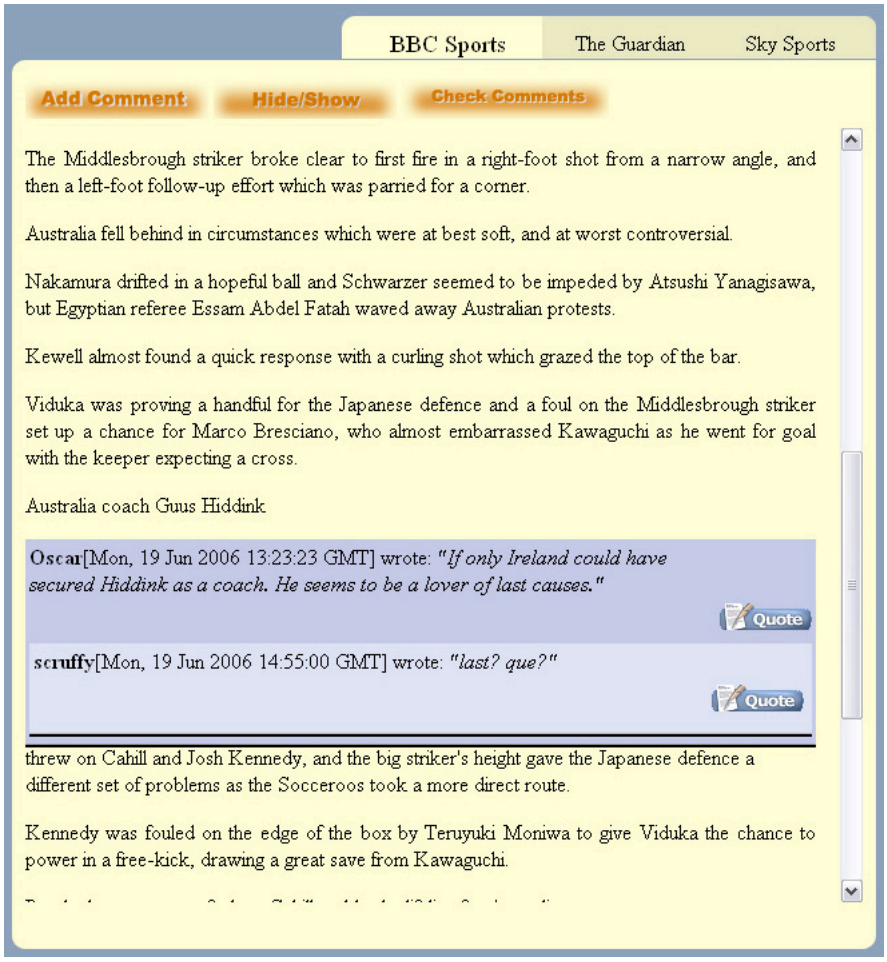


Figure 22: The reports panel with in-context annotations. The various buttons along the top allow for annotations to be hidden/revealed, as well as viewing of each of the different reports.

The alternating background colour for each post is used to signify the depth of the comment. If two replies are posted to the same parent, the same colour background is used. Also, a thick black ridge is used at the bottom of each comment depth as a visual aid.

In order to facilitate this interaction amongst users, several technologies were used in the system. In the following section each of the components of the system is presented, along with the system architecture.

3.2.1.1 Architecture

Since SportsAnno brings together information from different media sources and of different types, there is quite an extensive pipeline through which information must pass before being presented to the user. Figure 23 illustrates this pipeline. Information comes from two main sources before being gathered into a single match record; video recorded

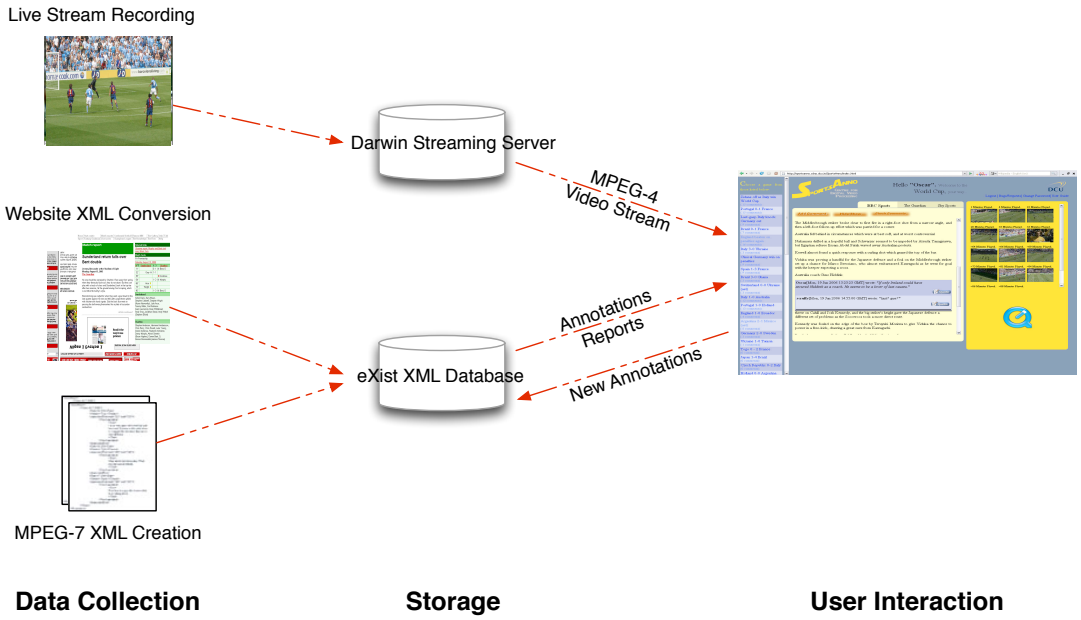


Figure 23: SportsAnno system architecture

from television and web reports taken from newspaper web-sites.

The initial video recording was made in MPEG-1 and later transcoded to MPEG-4, a step necessary to allow for streaming video playback through the Darwin Streaming Server[10]. Post-processing of the recorded video was done so as to remove all non-game footage such as studio discussions. This includes frames before and after the whistle. In this way the analysed video begins just before the initial whistle is blown and ends just after the final whistle. Each video was thus approximately 90 minutes in duration, deviations being due to extra-time and penalty shoot-outs.

Playback is shown through the Quicktime plug-in at the bottom right of the screen. It is possible to watch the entire match by clicking play on the player. Using the keyframes however will begin playback at the chosen event. The Darwin Streaming Server serves up the video in MPEG-4 format.

3.2.2 Summarising Sporting Events

Systems for the summarisation and browsing of sports video do exist (Liu and Zhang, 2005; Nemrava et al., 2008). None of these systems, however, present a written source of complementary information for the summarized video. Indeed much of the work in this field is on the continued automatic detection of highlights, players and events of interest within sports video.

[10]http://dss.macosforge.org/

While systems have been developed to enable browsing of event-detected video highlights, there are no systems which allow for this video to become the focal point of discussion. As has been said earlier, nearly all broadcast sports are accompanied by newspaper reports, blog entries or personal email correspondence. When in a public form, this additional written material can be used to offer a richer interpretation of the broadcast video. The emotive nature of sports means that this written material can be exploited to reveal valuable additional information. These resources also provide within themselves a means for discussion and interaction between users.

Video Segmentation: Once the video had been edited and cut to consist of just match footage, we ran a shot-boundary detection algorithm. The shot boundary detection algorithm used was the *Cut_detect* algorithm proposed by researchers within the Centre for Digital Video Processing (O'Toole et al., 1999). *Cut_detect* is a shot-cut detection algorithm for MPEG-1 video files. The approach is based on the quantification of frame-to-frame dissimilarity, implemented via the generation of metrics relating to both histograms and statistical moments for the colour components of each video image. Based on these descriptors, the algorithm invokes a threefold thresholding mechanism to quantify the significance of dissimilarity between frames, towards the detection of abrupt shot cuts in the video.

Since football video contains many hard cuts, the number of shots detected is very large while their duration can be very short. Each detected shot is assigned a confidence value based on how likely it is an event has occurred within the shot. Once events have been detected within the video, the shots are combined so as to provide segments that are of a more appropriate and usable length. The minimum length of a segment was chosen to be 15 seconds. The shots detected by *Cut_detect* are amalgamated into segments where all shots within the 15 second limit are concatenated to form a new segment as shown in Figure 24. If however the bounds of a shot containing an event overrun the 15 seconds limit, the amalgamated segment is increased so as to include all of the shot's event. Keyframe extraction is also performed with keyframes chosen as the middle frame of a segment.

Events are considered to occur when the event confidence value rises above a predefined static threshold and continue until this threshold is crossed again. This is shown in Figure 25. A description of the manner in which events were detected is beyond the scope of this thesis. All event detection was based on the work of Sadlier and O'Connor (Sadlier and O'Connor, 2005). The detection approach used was multi-modal and relied on both audio and visual information streams to determine confidence levels. Six Support Vector Machine classifiers where used which detected the presence of player close-ups, crowd shots, scoreboard changes, increased audio activity, playing field boundaries and increased visual activity.

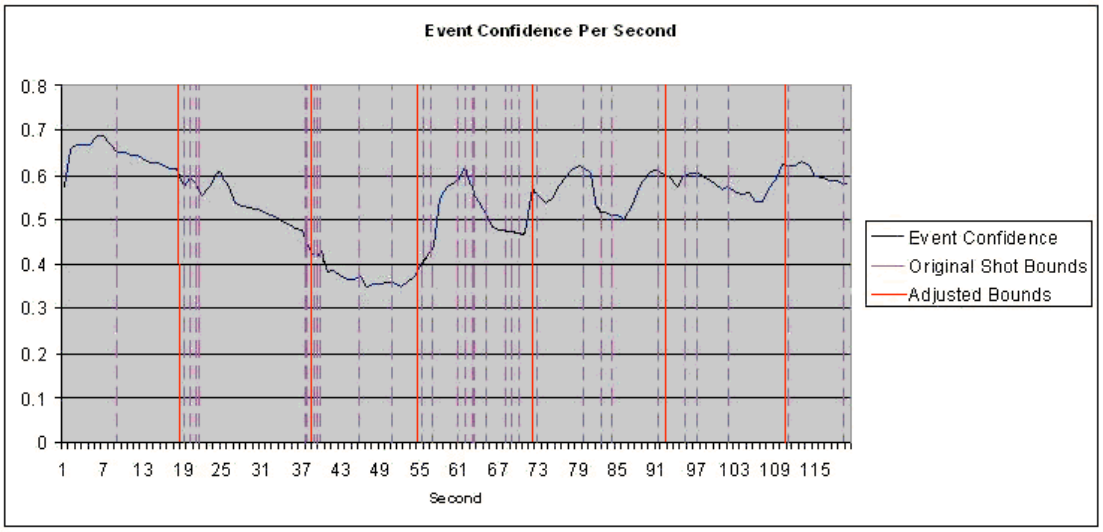


Figure 24: Creation of segment boundaries based on both the SVM classifiers and also the original shot boundaries.

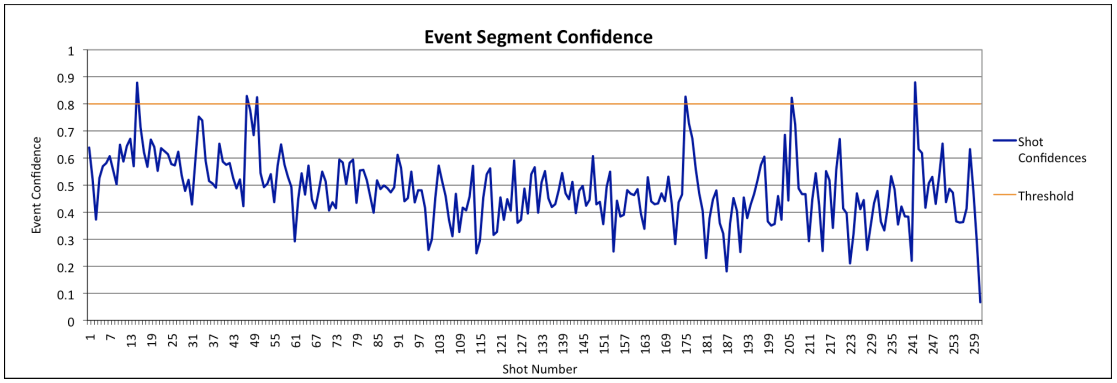


Figure 25: Events are detected whenever the confidence value raises above a static threshold. This threshold was set to 0.8

XML Document Storage: After finalising the segment boundaries, an MPEG-7 file was created which contains all the shot information including duration, start point and confidence of an event occurring during each segment (see Appendix C).

The second source of information required for each game is the match reports. Using a web parser, these reports were retrieved from the BBC Sport, Sky Sports and Guardian Unlimited web-sites. They were then transformed and stored as XML files. These three sites were chosen to give a cross-section of opinion. While the BBC and Guardian are less biased and brash in their coverage, Sky Sports was deliberately chosen as a site that would evoke more discussion due to its strong opinions.

Annotations were stored in separate files from the original report so as not to alter the original document. This was done so as to easily identify the insertion point for comments regardless of the number of comments already made. The benefits of storing

annotations separately from the original documents have been noted previously (Bottoni et al., 2004; Kahan and Koivunen, 2001; Ovsiannikov et al., 1999). An XML structure was specifically created so as to maintain the thread structure of the annotations. Each annotation has within its record the name of the author, time of creation, quoted text and its content. As stated, users may create comments anchored on any sentence or paragraph within the original reports. Replies to annotations are considered to be focussed the entire text of their parent annotation.

In order to organize the different files required for each game, a master file was created which links all the XML files of a match. This is the cross-reference file that contains the names of the report files, the annotation file and the MPEG-7 file of the match.

SportsAnno was built using an XML backbone so as to enable easy integration of existing standards whilst also providing easy extensibility. All data files required by the system are stored within an eXist XML database[11], a freely available open-source project. The eXist database provide all the required functionality of a database for the storage and query of XML documents.

3.2.2.1 Usage Study

SportsAnno was closed in nature and so the user base consisted of people either directly known by the authors or know by a direct colleague of the authors. 70 people registered with the system, the dates of registration varying greatly from before the competition started to within the last week of the FIFA World Cup. All games were made available to all users however, so even those who registered late could browse and comment on any match including those played before registering. 25 of the registered users were researchers within the group who had experience of annotation systems. A further 12 came from associated research institutes who would again have had experience with annotation. The rest of the user community was made up from friends of registered users.

Almost 83 hours of video data was recorded over the duration of the competition, consisting of 54 matches. This was accompanied by 162 newspaper reports. Not all matches were available for recording due to scheduling conflicts on television and one game (Serbia Ivory Coast) was lost due to a recording error. The remaining matches were all fully indexed and processed for event detection.

From the 70 people who registered, 24 made no further visit to the system. Of the remaining 46, 24 were active browsers viewing the comments left by others but not

[11] http://exist.sourceforge.net/

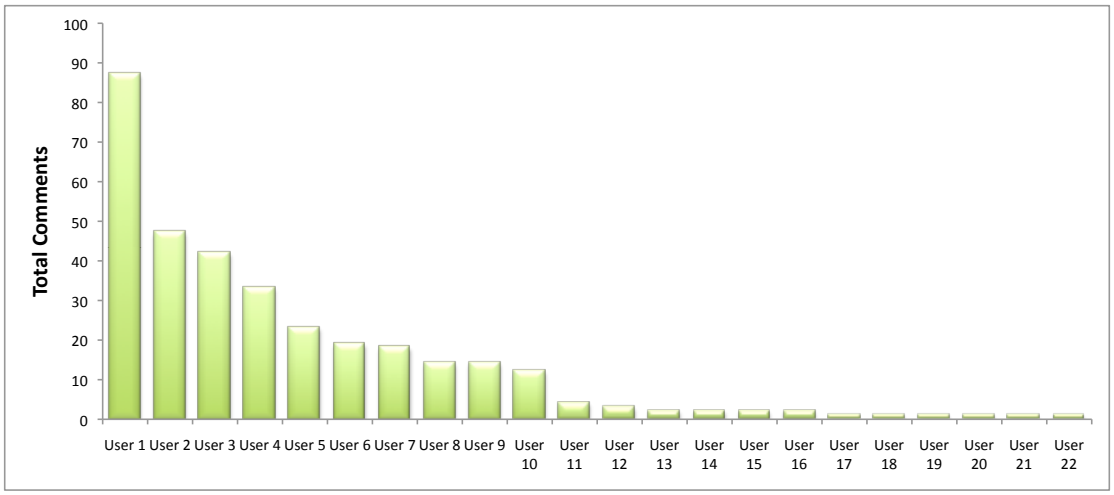


Figure 26: Number of comments created per-user within SportsAnno

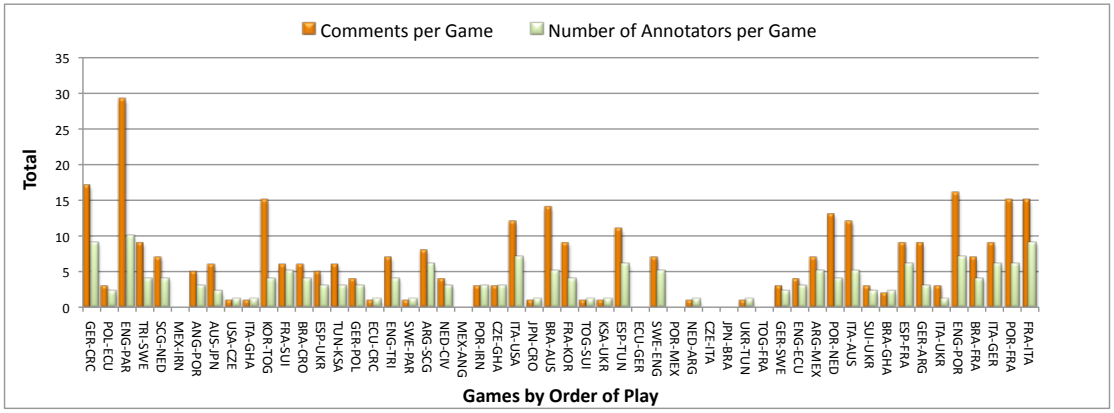


Figure 27: Number of Comments and Annotators per-match within SportsAnno

contributing themselves. 22 users made comments and took part in discussions about particular events within each match. The number of comments made by users varied greatly as seen in Figure 26. These comments are made up of both replies to previous posts and original postings. No differentiation is considered here. Figure 27 shows how many of these users commented on each match during the competition. The 22 users were not only those users who had had past experience of annotation systems.

It is clear that the first England game against Paraguay was particularly well commented. This is not a surprising result as the hype surrounding the England squad within the media of both the UK and Ireland generated lots of talking points. Only 6 of the 54 recorded games received no comments. Again, these games involved teams that would have little following within the registered audience, the only surprise perhaps being the Brazil-Japan fixture.

The ratio of commentators to comments shows that commenting is a useful way in

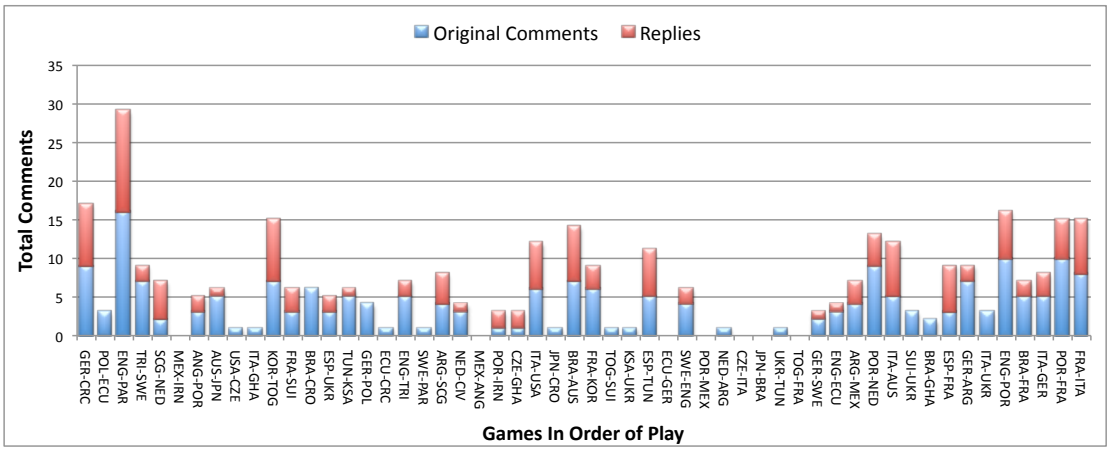


Figure 28: Comments per-match divided into replies and direct in-context comments

Table 3: Distribution of comments within the SportsAnno corpus

	Comments Per Game	*Original Posts*	*Replies*	*Comments After 3 days*	*Commentators Per Game*	*Thread Depth*
Average	6.1	3.7593	2.35185	0.72222	3.25926	2.06
Std. Dev.	5.84603	3.26754	2.9084	0.14973	2.49668	1.41976
Max.	29	16	13	6	10	6
Total	330	203	127	39		

which to generate discussion within a group. Within games with more than one user's comments, it can be seen that it is not just new comments which are added but instead replies to the comments already left. Figure 28 shows the number of replies per game, broken down into original comments (i.e. comments which are not in reply to another comment) and replies. It can be seen that where original comments are attributed to more than one user, the number of replies versus original comments is high.

One of the possible reasons for users not creating more replies to comments was the lack of a notification system which could notify users when a comment they had made was replied to. In this way, a user's attention would have been more readily drawn to the specific reply.

The time between first posting the match to the web-site and the last comments on a game being made was also recorded. Due to the type of data being presented, it is not altogether surprising that the number of comments made on a match fell dramatically 3 days after its first posting. Some games proved exceptional, mainly those involving teams that stayed in the competition for longer. Users did post comments on earlier performances involving teams such as Germany (the hosts) and France (the current champions of the World Cup) but in general, comments were of a more immediate and transient nature.

Table 3 shows the number of comments made and who was making these comments. The maximum figures for comments in the first three columns all correspond to the England-Paraguay match. This was England's first game of the competition and so generated a lot of discussion. We can see that while there are more original comments than replies, there is on average at least 2 replies to each post. The deviation is due to the existence of both highly commented games and those which received no real discussion. As mentioned before, replies are more prevalent for games where more than one person has created original comments.

The number of days after which the game was commented on was affected most strongly by the advent of weekends (during which very few comments were made) and the amount of time between the recording of the matches and when they were made available on the SportsAnno web-site. This time varied from same day to 2 days after the recording date. It is also thought that lack of a notification system prevented discussion from having an average life-span of greater than 3 days, as mentioned earlier.

3.2.2.2 Observations and Reactions

SportsAnno gave us our first experience with creating a truly multimedia browsing environment in which discussion could flourish. Although we did not have a large user group, the system did prove to us that users relish the opportunity to become more actively involved in the publication process. This system also showed that the inclusion of comments does provide additional information to the corpus, information which is of use to the community of users as a whole. For more information on the level and distribution of annotations, the reader is referred to (Lanagan and Smeaton, 2007).

After the experiment's completion, we interviewed a cross-section of users for their opinions on the system. These users varied from highly-active users of the system who made many comments on different matches, to those who used the system less often and more passively. Users were interviewed informally face-to-face as most were known to us and easily contactable. The suggestions made to us are aggregated below.

Keyframe Vagueness: The most prevalent complaint was that the keyframes chosen to represent events and displayed on the right of the interface bore no real relation to the events. As a result, the keyframe itself was of little use in knowing what would be displayed when it was clicked. This meant that the caption above the keyframe showing the time of the corresponding event was very important, providing the only real

indication of what the event was. One suggestion was to present a summary of the event in words, or simply a tag such as 'Goal', 'Foul' etc.
The lack of anything to distinguish the keyframes was a problem which arose from automatic keyframe selection. As stated, this was simply the middle frame from within an event. Improvement to event boundary detection was shown to improve keyframe selection in Annoby (see below).

Accessing New Comments: While the indication of new comments on the navigational side-bar was said to be useful, no method of showing just new comments within the actual reports was present. This resulted in having to search through all comments on a game in each report, until a new comment was located. This was deemed acceptable due to the small number of comments, but could be frustrating were the scale of the system to rise.

Another concern was an **inability to filter comments** in any meaningful way, such as based on time, author etc. Again this would cause more problems were the scale of the system to increase. The fact that annotations were presented in-context was beneficial to users, although the ability to create separate threads was requested. While this feature was desirable, the premise of the experiments behind the development of SportsAnno is the creation of in-context annotations. This would be worth noting when creating an annotation system, but we feel that it is outside the scope of this thesis and so no further consideration is given to it.

Other suggestions included **direct annotation of the video**, allowing users to define where in the video to place annotations. This however is similar to the creation of separate threads of annotation not contained within reports, and so this feature was again given no real consideration. An area showing **general information on the match** being viewed is also desirable. Many people had watched games live before using the system and so video was not as central to their experience. The entire footage of the game was present and available for viewing, but navigation was not fully implemented. This was mainly down to the fact that the video was aimed at showing the highlights already detected. Easier navigation of the entire video remained a request.

One other observation was the distribution of the **locations** of comments/annotations. The largest percentage of comments appear on the first external news report about a match, with nearly all being confined to the first two. This is perhaps a consequence of all three reports being re-wordings of the same events. We shall show that it is important to randomise which report is presented to the user first so as to create a more even distribution of comments across the reports.

Figure 29: The Annoby main interface

3.2.3 Annoby

After analysis of the SportsAnno annotation data-set and system, we took the opportunity to improve and build upon the experience gained in this experiment. As stated, some of the major drawbacks to annotation creation and discussion was through unforeseen weaknesses of the functionality of SportsAnno. With our next system, Annoby, we aimed to fix the most significant of these and introduce new possibilities in user interaction.

As with SportsAnno, Annoby is a closed system allowing access to registered users only. When first accessing the system, a short video highlighting all the features present in Annoby is shown to the user. This is done because many of the features are not commonly found in web-sites, and would have gone unnoticed to the detriment of both users and system. The list of games presented to users in the SportsAnno system was removed as it was deemed unnecessary, it also would have required additional loading of a separate interface which was considered a disadvantage. Instead users are presented with the main interface (Figure 29) through which all interaction takes place.

The most recently uploaded game is presented by default with the assumption that this game will be the current point-of-focus for users, and therefore where the most annotation will have taken place since a user last logged in. On the left of the screen the list of games is also available for easy navigation between matches; to the right of the screen, a list of all the commentators on the current game. Both of these lists are sortable, allowing users to organise them by comments, match/comment date or alphabetically. This allows the focus to shift between currency and activity. The main focus of the interface is now the centre of the screen, with attention no longer split between keyframes and reports. Instead, keyframes are presented in-context within the reports. Video playback remains to the right of the reports as it was in SportsAnno.

The details of the match are shown above the main reports in the centre of the screen (Fig. 30). The tag-line for the game is taken from the headline of The Irish Times report, providing a one sentence overview of the match. Also included are the location and date of the match. Next to these, comment statistics provide a quick idea of how much conversation has been taking place about the game. By providing an individual thread count as well as the number of annotations and commentators, we have an estimate of the depth of conversation. A game with several comments and few threads helps to show that conversation has been focussed on a few key points. It is also more likely that this conversation will yield interesting information which was not within the original reports.

Figure 30: Match and commenting details

Commenting: As was the case with SportsAnno, the purpose of Annoby is to create a corpus of annotations/comments on which to test the algorithms developed in this thesis. We have tried to make commenting even easier in this second system implementation, removing many of the issues described by users of SportsAnno. Comments are created by highlighting a portion of text within the reports. When the mouse button is released, a pop-up appears which inserts a comment box into the report.

Comments are no longer as intrusive as was the case in SportsAnno. SportsAnno

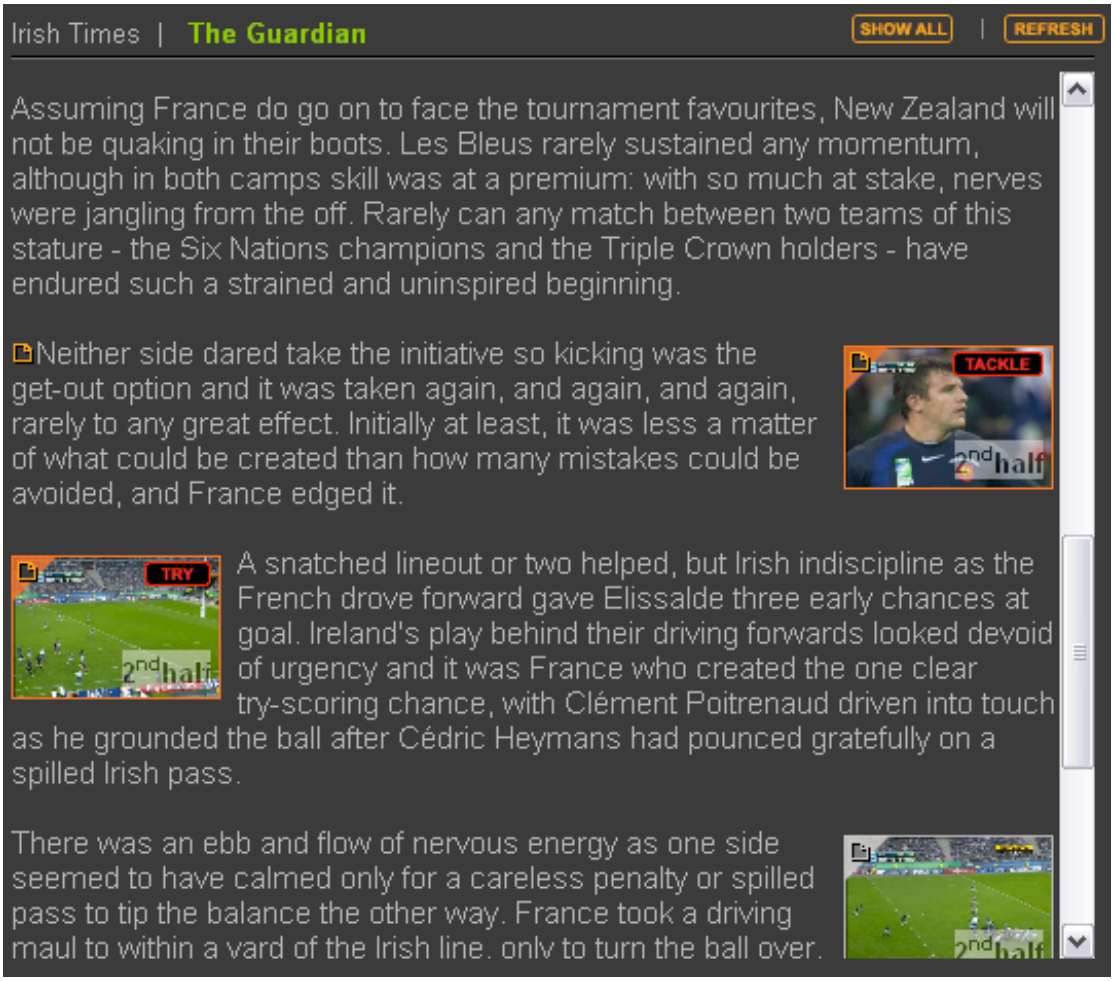


Figure 31: The Reports Panel with in-context annotations. The various buttons along the top allow for annotations to be hidden/revealed, as well as viewing of each of the different reports.

only allowed for hiding and showing of all comments at once, something which could be disorientating when large numbers of comments had been made. While the option to show all comments is still present via the "Show All" button, single comments may be shown by clicking on the small orange annotation symbol, an example of which may be seen before the second paragraph in Figure 31. Once the comment thread is revealed, replies to a comment may be added by clicking the 'comment' button, or the comment box is hidden by clicking on the small grey arrow in the top corner.

It is also possible to highlight the comments of a particular commentator, allowing users to easily find the new comments or comments which most interest them. By clicking the small 'Show' button beside each commentator on the top right of the screen, all threads containing comments by that person in either report are revealed. The comments themselves are then highlighted as shown in Figure 33. Clicking on the 'Show' button also changes the text to 'Hide', enabling the minimising of annotations as well as highlight removal. This ability to highlight the comments of a particular users

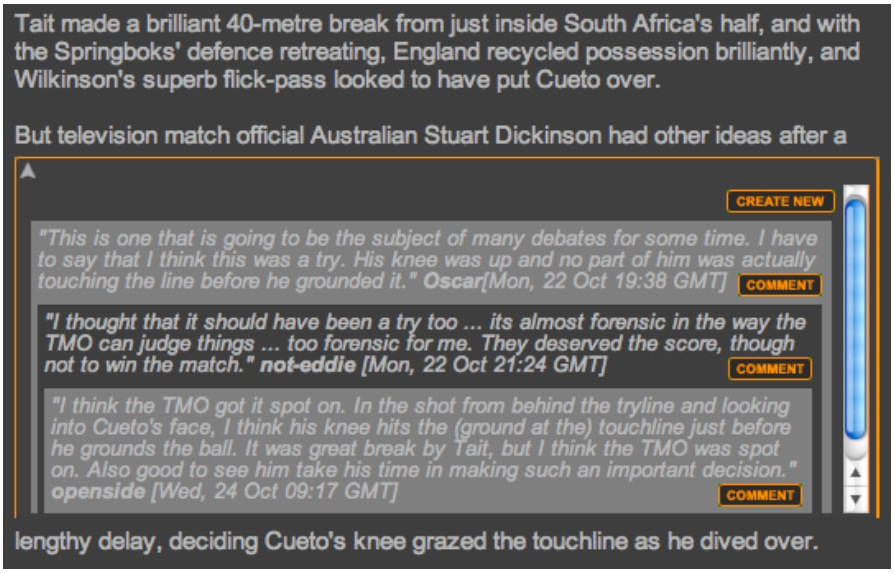


Figure 32: A thread of comments

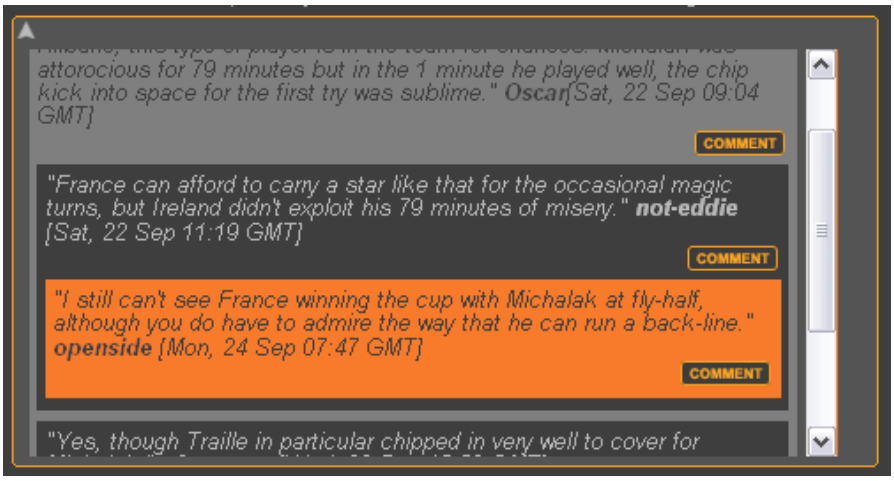


Figure 33: Comments by a particular user may be highlighted so as to be easier to find.

is important as it helps to provide a means of seeing the general charectoristics of the annotator in question. Being able to gain a sense of an annotators general stance and viewpoint on particular topics has been shown to influence the way in which other react and interact with them (Wolfe, 2000).

With the combination of keyframes and reports, the ability to annotate a keyframe and implicitly the event which it represents was introduced. An un-annotated keyframe is present at the bottom of Figure 31 showing a grey border. A grey border is used along with a grey annotation symbol in the top left corner to signify the absence of any annotations, but the opportunity to create annotations remains. In order to provide consistency, annotations on both event keyframes and free text are presented in orange.

Another important feature of the keyframes in the semi-transparent indication of time in the form of '1st half'/'2nd half'. With the constant stoppages of play in rugby, it was even harder to give an accurate estimate of the timing of events than was the case with SportsAnno. More important still is the use of tags on each important event within the video, allowing users to see at a glance what the keyframe represents. These tags where added manually as keyword annotations to the MPEG-7 files of each video.

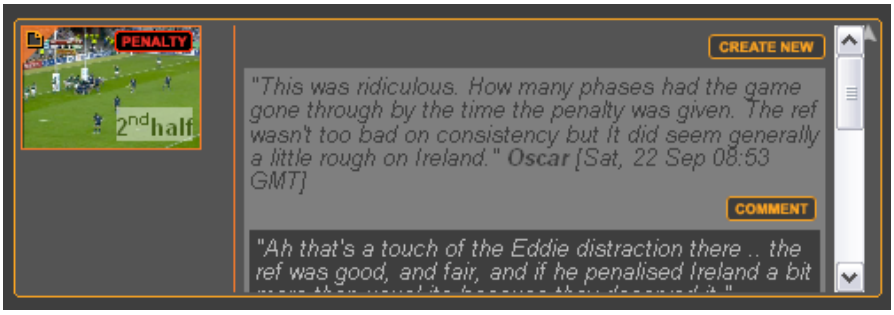


Figure 34: Comments may be made directly on the keyframes representing an event within the video.

As with the text annotations, a small orange symbol is present on an annotated image along with the border coloration. Clicking on this symbol reveals the annotation thread (Figure 34). In all other respects the annotations are identical to those made directly on the text so as to create the aforementioned consistency. Clicking on the image directly however will result in the playback of the corresponding event within the video.

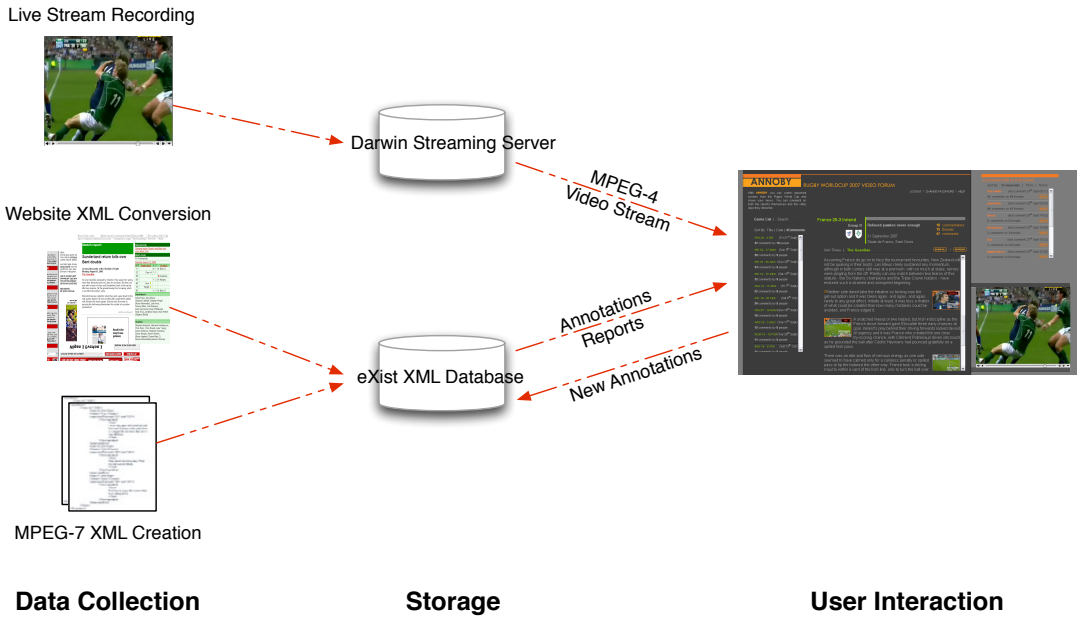


Figure 35: Annoby system architecture

3.2.3.1 Architecture

The Annoby system (Lanagan and Smeaton, 2009) was designed to give users an even more interactive experience of the events of the Rugby World Cup 2007. While much of the back-end architecture used was the same as SportsAnno, many significant improvements were made. These improvements focussed on the user interface (UI) but included an improved adaptation of the event-detection algorithm used for SportsAnno which had built on the work of Sadlier and O'Connor (2005).

The capture and storage of both video and text sources remained identical to those in SportsAnno, as illustrated by Figure 35. With Annoby we did however decide to scrape just two newspaper reports instead of the three used in SportsAnno. One of these came from the same Guardian Unlimited[12] source used in SportsAnno and the other from a new source, the Irish Times[13]. These two sources were chosen for the differing viewpoints and perspectives they would present. Both Ireland and England were present for the Rugby World Cup 2007 so these British and Irish publications provide a natural bias to their analysis. The number of sources was reduced to two as it was noted that in SportsAnno, the vast amount of annotations appeared on just 2 of the reports.

Event Boundary Detection: We used the same *Cut_detect* algorithm proposed by researchers within our research centre (O'Toole et al., 1999) as had been used to establish the shot boundaries in SportsAnno. Due to the insufficiently accurate detection of event boundaries using the SportsAnno extension however, we changed the way in which events were bounded.

Similar to live football video, live rugby video contains many hard cuts creating shots which vary drastically in duration. Again, initial shot boundaries are taken as the skeleton onto which the new event segments are fitted. Once shot boundaries have been detected, we calculate the per-second confidence values for the event boundaries of the entire video. These confidence values are based on the same six SVM classifiers used in SportsAnno. We then use these per-second confidence values to calculate the highest-valued event segments within the video. This time we use a 20 second event-window as it proved better at combining highly rated shots which belong to the same event segment. To calculate the event boundaries we proceed as follows:

1. Sort the per-second confidence values in descending order.
2. Find the ten highest per-second confidence values within the video. If the threshold is reached, take high-confidence values found so far.
3. Extend a window of 20 seconds around each high per-second confidence value,

[12] http://sport.guardian.co.uk/
[13] http://www.irishtimes.com/sports/

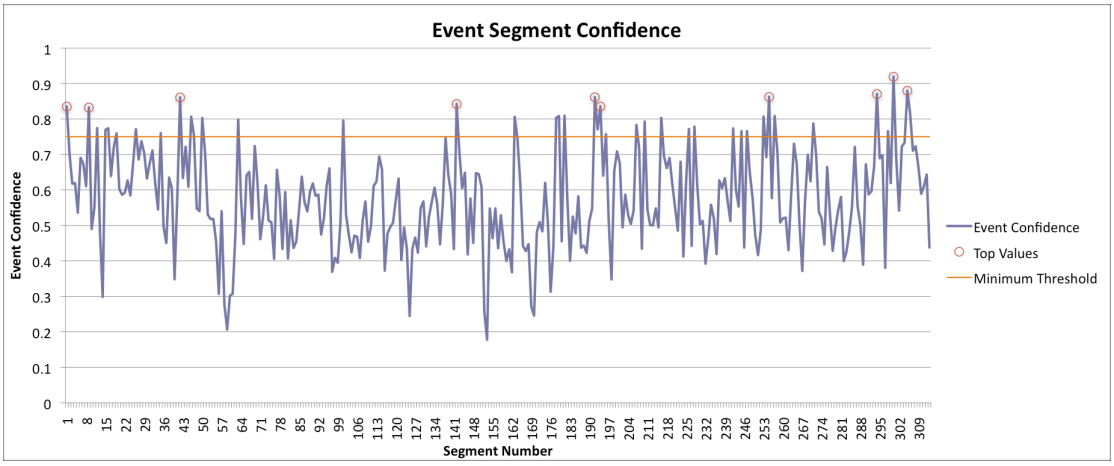


Figure 36: Only the top 10 most exciting events are returned, and only if these 10 events have a confidence higher than the minimum threshold. This threshold was set to 0.75

centered on the highest confidence value.

4. If any of the windows overlap, combine the windows into one. This is done so as to amalgamate high-confidence shots which belong to the same event segment.

5. Match up the high-confidence windows to the shot boundaries already detected. Expand the windows so any shot within the window is complete contained.

6. Windows now contain the top 10 events for the video.

7. Check if the threshold has been reached for finding high-confidence shots. If so exit.

8. If any windows have been amalgamated, go to Step 2.

One of the most significant changes made to the event detection algorithm in Annoby is the use of an expanded event window and dynamic threshold. There were occasions when the average confidence level for a video was very high resulting in lots of high-confidence shots. In SportsAnno this meant that a large number of keyframes were displayed, further obfuscating the meaning of each keyframe. With Annoby we chose to restrict the number of events returned for a video to the top 10. Some games did not have 10 highly exciting events and so the dynamic threshold was combined with a lower bound so as not to return meaningless segments.

The expansion of the event window was as a result of initial testing. The nature of rugby video is slightly different to football video due to the drastically different rules of the game. Rugby Union[14] is a game in which there is much stopping and starting, not unlike American Football. As a result the shot detection algorithm can have more

[14]http://en.wikipedia.org/wiki/Rugby_union

challenges. By expanding the event-window, the chance of high-confidence related shots being contained in the same event is increased.

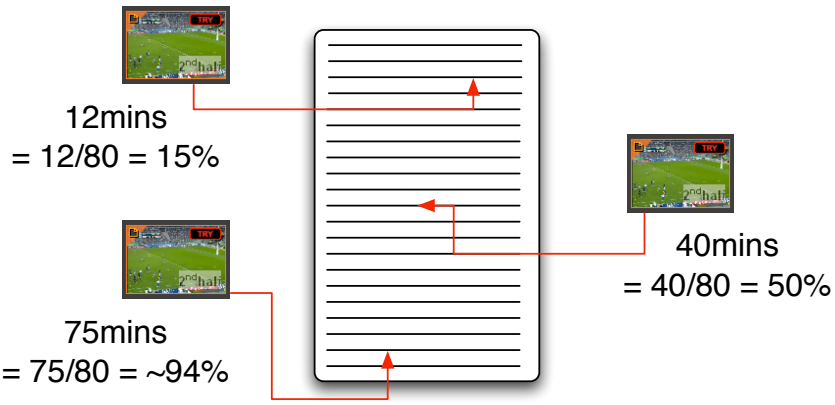


Figure 37: Keyframes are inserted at the same percentage-region of the written reports as they appear within the video

As stated, the focus of attention within the Annoby interface is on the reports pane (Figure 31) which now holds both keyframes and reports. The splitting of keyframes and reports within the SportsAnno interface meant that connection between video and written reports was diluted. With Annoby, keyframes of the most significant events were presented in-line within the report so as to make the connection more explicit. The position of keyframes alternated between left and right, each keyframe being presented at approximately the same offset into the report as the percentage time into the match. Sports reports are theoretically written with the first paragraph summarising the entire game, events are then presented in chronological order providing an outline of the match as a whole (Andrews, 2005). Using this fact and without any semantic analysis, we can present video events in the region of their corresponding text description.

3.2.3.2 Usage Study

As with SportsAnno, the registration period for Annoby ranged from before the start of the competition to any time during it. All games were made available to all users and so even those who registered late could browse and comment on any match including those played before registration. Of the 89 people registered with the system, 25 were researchers within the group who have had had experience of annotation systems. 14 of the registered users for Annoby had also registered and actively used the SportsAnno system. Again, with Annoby being a closed system the user base was drawn from a community of people known to the authors or were a friend of a friend.

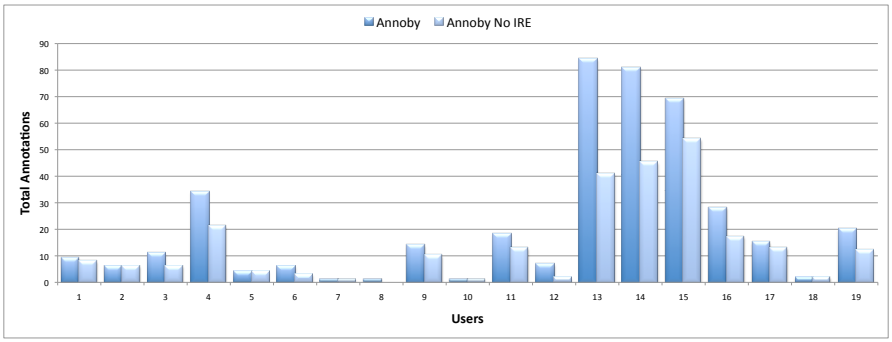
48 matches were recorded creating almost 66 hours of video data over the duration of the competition. This was accompanied by 96 newspaper reports, 2 reports per-game.

These matches and reports were fully parsed and indexed for use in the system. Of the 89 people who registered, 50 made no further visit to the system. Of the remaining 39, 20 were active browsers returning to the system at least twice and viewing the comments left by others, but not contributing themselves. The remaining 19 users took part in active discussion, creating the annotation corpus for the Annoby system. These users were not solely those with past experience of annotation systems.

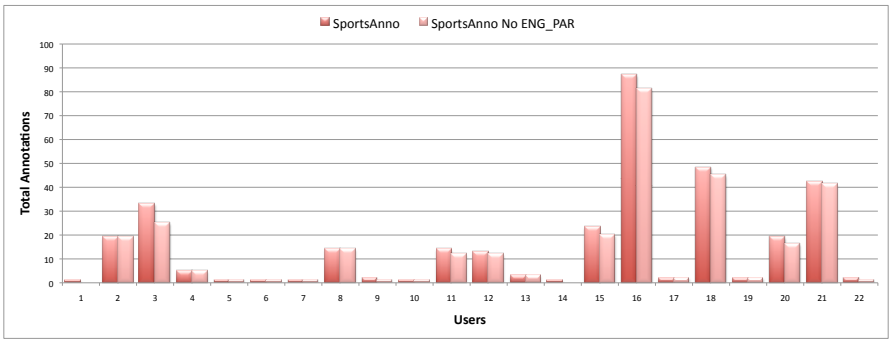
The total number of annotations made on the Annoby system was 411, slightly higher than the 338 made in the SportsAnno system. (As noted later, the effect of Ireland's presence in the Rugby World Cup 2007 should not be ignored, these games making up 152 annotations. The most highly commented games in SportsAnno was the first England game with 29 annotations.) The distributions of these comments across the reports however is roughly similar. SportsAnno did not randomised the initial report shown to users when a game was viewed; it was always the BBC reports which received 65.66% of all annotations. With Annoby the number of reports was reduced to 2 and an attempt to randomise the default report was made, however this randomisation was not implemented from the start. As a result we note that the Irish Times reports received 67.88% of annotations. While this is approximately the same as the BBC reports of SportsAnno, it is difficult to draw concrete conclusions due to the late implementation of randomisation. It would appear that when presenting more than one match report (each of which is essential the same summary but from a different stand-point) for viewing and annotation, randomisation of which report is shown by default is important. Without this, any more than one report is redundant as users appear to only read (and therefore predominantly annotate and discuss) the first report they are presented with.

The number of comments made by users varied greatly as seen in Figure 38. These comments are made up of both replies to previous posts and original postings. Though no differentiation is made between these in either Figure 38 or Figure 39, the information has been recorded in order to test the algorithms presented in Chapter 4. In the analysis that follows, we present statistics for both Annoby and SportsAnno systems. While we are aware that there are several factors which need to be taken into account when making comparisons between the systems (user familiarity with annotation systems; incident levels within the respective sports; overall viewership figures for both sports), the underlying purpose of both systems is identical.

An important consideration which must be made is the presence of Ireland within the Rugby World Cup 2007. The effect of this can not be underestimated since the number of annotations received by each of the Irish games is far above the average number of annotations per game. In the Football World Cup of SportsAnno, Ireland were not present, however England were and acted as a proxy or substitute for focussed attention. This may be seen from the number of annotations the first England game (England Vs. Paraguay) received. Again this is far above average being the first game



(a) Number of annotations per user of Annoby with/without the inclusion of Ireland's matches



(b) Number of annotations per user of SportsAnno with/without England Vs. Paraguay

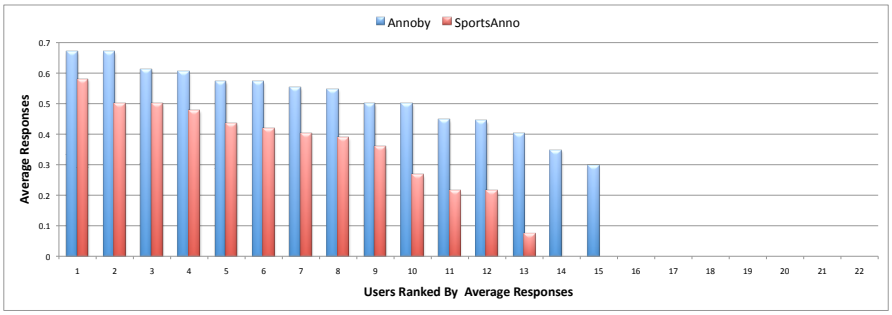
Figure 38: The average number of annotations per user of our systems.

England played after a highly talked about and controversial team selection. For these reasons we have presented each of our results and analysis in two ways; we include all Irish games and the highly commented England vs. Paraguay game in our analysis and then exclude these 5 games.

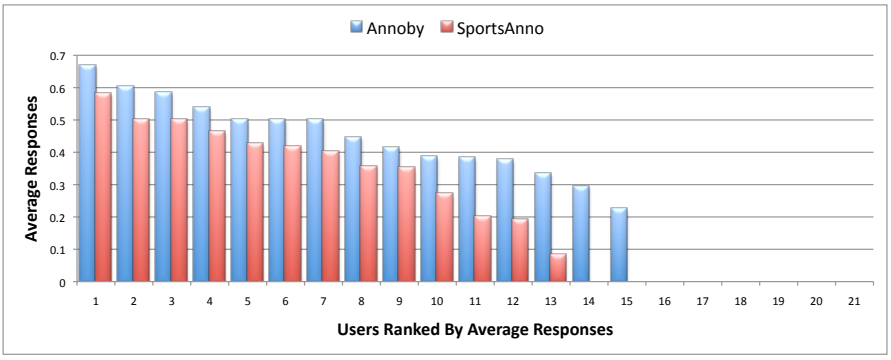
The effect of removing these games can be seen in Figures 38(a) and 38(b) where the total number of comments per user is displayed. By removing the Irish games we see a dramatic decrease in the number of comments made by the most active users. The number of users who posted comments on each system is also reduced, showing that some users commented solely on these 5 games. Two fewer users are present in SportsAnno having created a single comment on the England vs. Paraguay game. In contrast although we have removed 4 games from the Annoby system by removing the games involving Ireland, just 1 less user is present in the Annoby statistics of Figure 38(a).

3.2.4 Comparison of SportsAnno and Annoby Usage

Figure 39 shows the average number of replies received to annotations made by users of both the Annoby and SportsAnno systems. It is clear that the number of replies received by postings in the Annoby system is on average higher. As noted above, the



(a) Average responses per user including Ireland's matches, and England Vs. Paraguay



(b) Average responses per user excluding Ireland's matches, and England Vs. Paraguay.

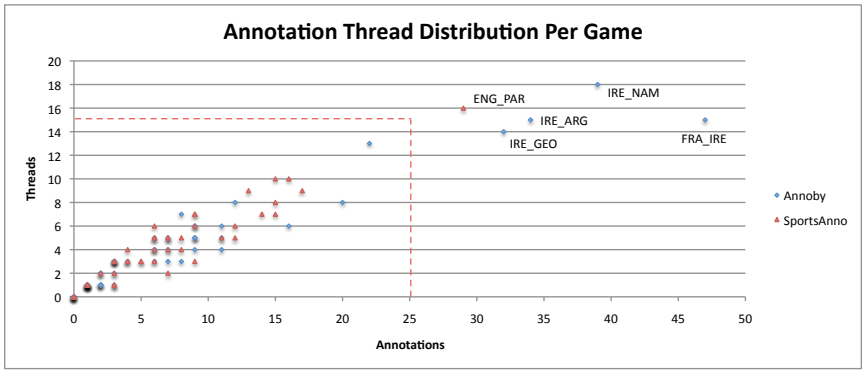
Figure 39: The average number of replies a user's annotations received.

Table 4: Comparison of average annotations per-game received by Annoby and SportsAnno

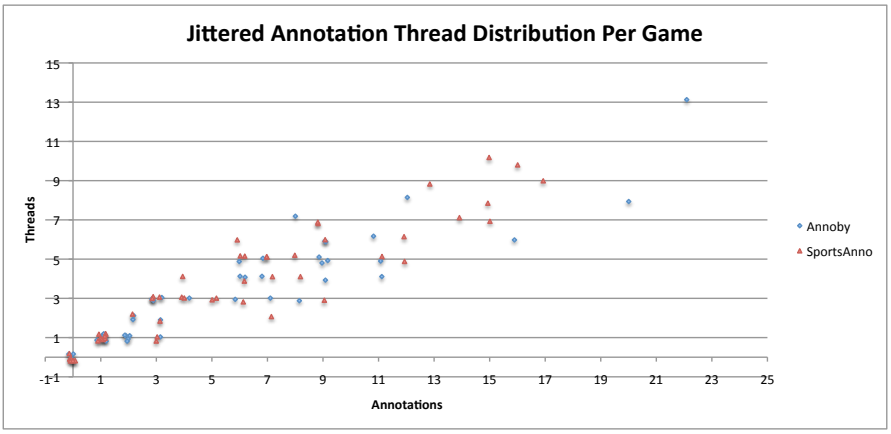
Annotations	*SportsAnno*	*Annoby*
Total	8.58	6.12
Text	6.15	6.12
Image	2.43	-
Without Ireland/England vs. Paraguay	5.89	5.68

total number of annotations by the top ranked users (rank by annotation creation) is greatly reduced when Ireland's games are not considered. This leads us to believe that there is a great amount of conversation being held around these games, mostly by highly-active users. The average responses to comments in Annoby however remains greater even with the exclusion of Irish games. If Ireland were the only reason for users being more interactive and conversational, we would not expect this to be the case. This fact, as well as the answers to a survey carried out, lead us to believe that the Annoby system made conversation-building easier and more engaging.

The average number of text annotations per game in Annoby (6.146) was similar to that of SportsAnno (6.115), however games in Annoby also received an average of 2.427 image annotations. Figure 40 shows the two distributions from Annoby and SportsAnno



(a) Correlation: SportsAnno(0.953) Annoby(0.954)



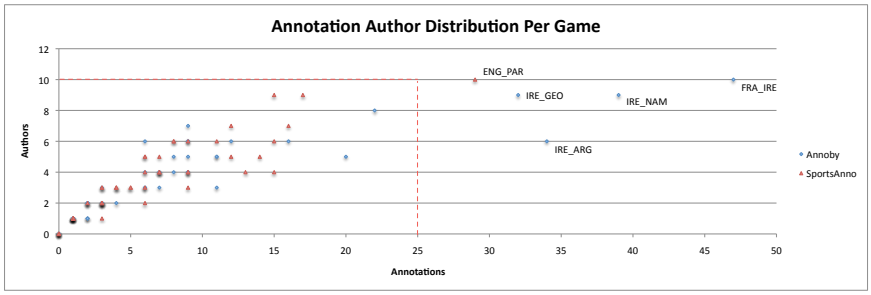
(b) Correlation: SportsAnno(0.999) Annoby(0.920)

Figure 40: Comparison of annotation thread distributions (a) With and (b) Without Ireland and England Vs. Paraguay

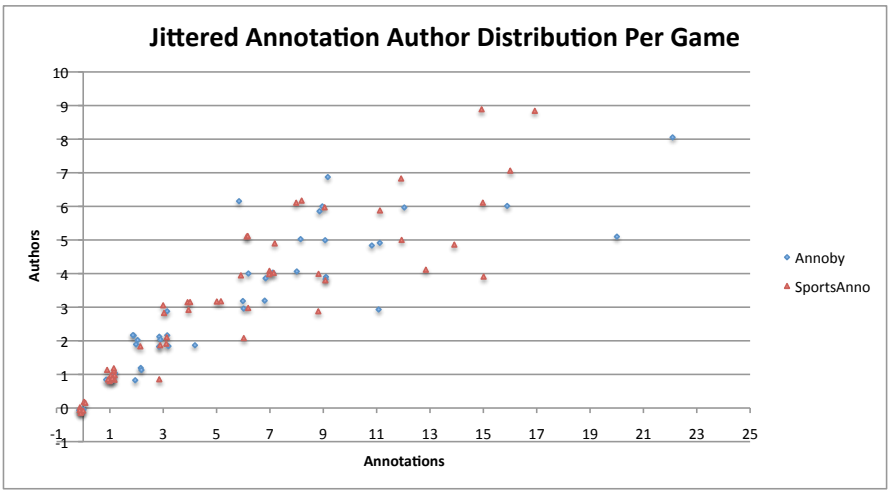
with and without the outlier games of Ireland and England vs. Paraguay. (The data-points in Figure 40(b) have been *jittered*[15] so as to allow a clearer visualisation of the data.) The significant impact of these games may be seen in the fact that without them, the average total annotations per game for SportsAnno (5.679) and Annoby (5.886) are almost identical. From these figures it would appear that the introduction of keyframe (and implicitly video) annotation does not in fact increase the average number of user annotations, but instead replaces an equal number of text annotations.

The number of users creating comments per game within the Annoby and SportsAnno systems may be seen in Figure 41. We can see that there is almost perfect correlation between the number of users and number of annotations created in the SportsAnno system, especially with the removal of the first England game as in Figure 41(b). While correlation is still strong in Annoby, the weakened correlation echoes the observation of Figure 39; increased responses to comments made by users in the Annoby system show

[15] Jittering is a process by which a small positive/negative number is added to the value of data-points. By doing so, the distribution of values may be more easily observed. In the example in Figure 40(b), we are able to see the two distinct plots.



(a) Correlation: SportsAnno(0.900) Annoby(0.854)



(b) Correlation: SportsAnno(0.999) Annoby(0.858)

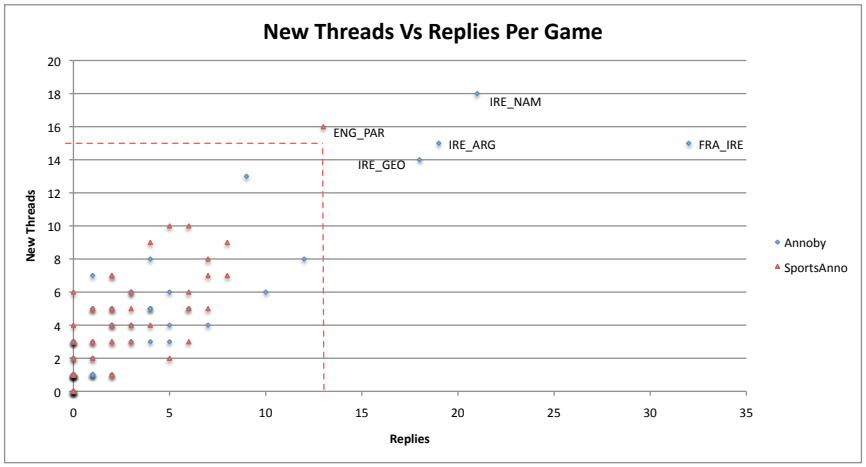
Figure 41: Comparison of annotation author distributions (a) with and (b) without Ireland and England Vs. Paraguay

that while semi-direct video annotation does not increase the number of annotations created by users, it does seem to increase the conversation and interaction of users.

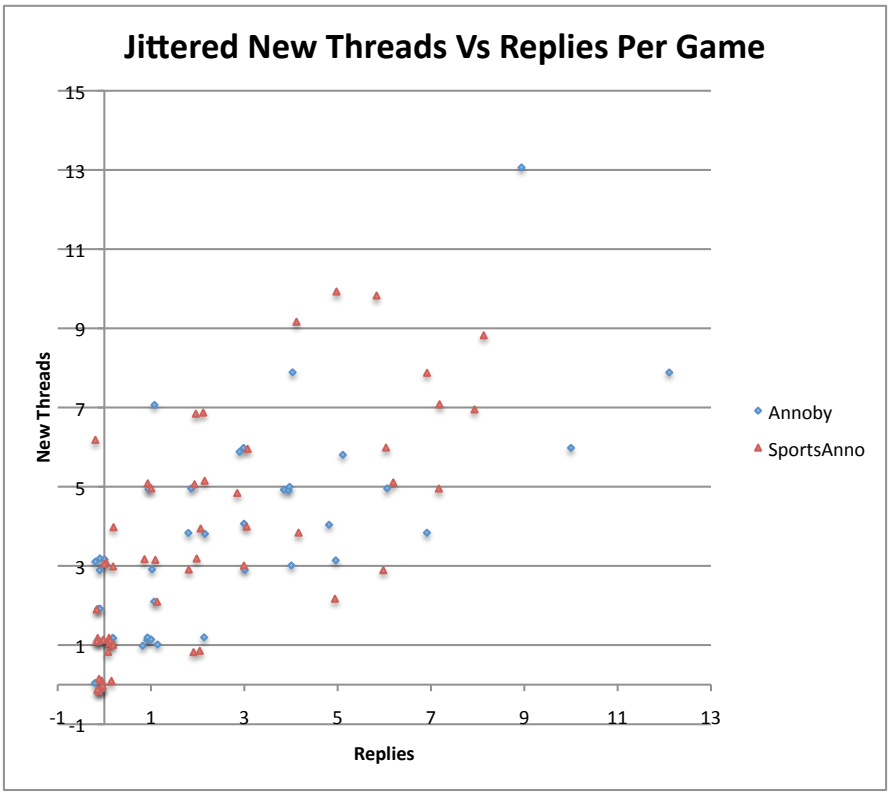
From the distributions of annotation types shown in Figure 42 and Figure 43 we can see that the behavior of annotation does not seem to be affected by the type of sports video being annotated. Indeed once we remove the 5 most highly commented games as shown in Figure 42(a), the distribution of replies to new threads created is highly similar. This fact seems to be true both on a game basis (as in Figure 42) and also on a user basis (as in Figure 43).

3.2.4.1 Observations and Reactions

As with SportsAnno, after completion of the experiments we asked a random selection of 8 users about their experiences with the Annoby system. This was done in the form of a questionnaire, a copy of which may be found in Appendix A. The experiences of users with the new system (which took into account the suggestions made by users of the SportsAnno system) seem to on the whole have been good. Again the users surveyed



(a) Correlation: SportsAnno(0.791332538) Annoby(0.875767168)

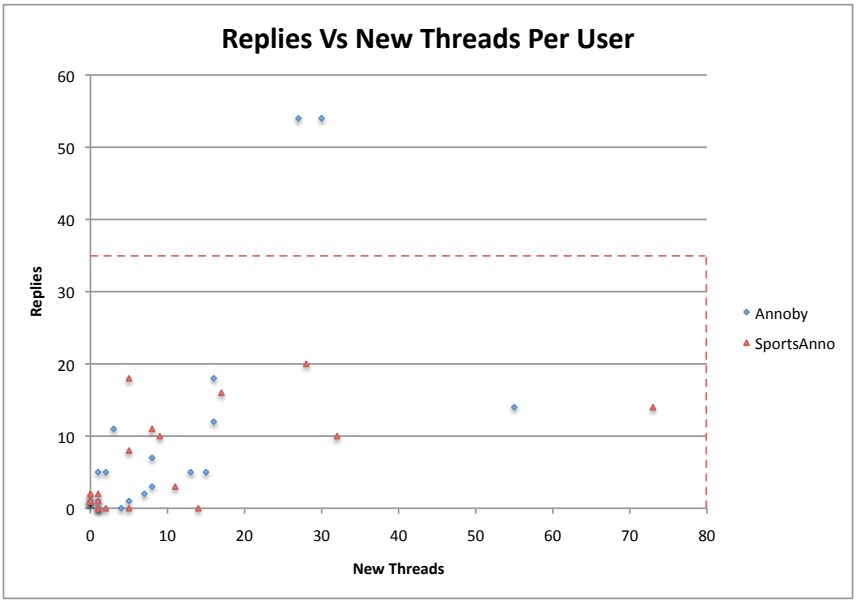


(b) Correlation: SportsAnno(0.716705863) Annoby(0.722192758)

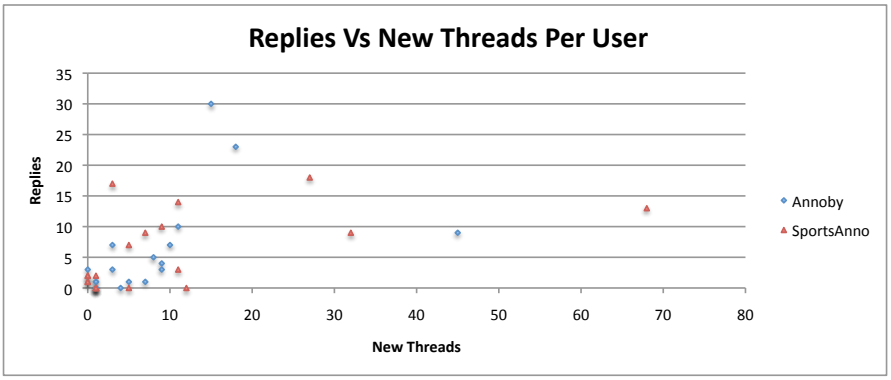
Figure 42: The number of new threads started vs. replies to threads per match (a) with and (b) without Ireland and England Vs. Paraguay

ranged from the most active to those who spent more time browsing and reading the comments of others.

Initial assumptions regarding the usage of the system were disproved; it had been assumed that users would take advantage of the recorded matches to catch-up on and summarise the matches which they had not seen. In fact those surveyed preferred to browse the games that they had already seen, rarely bothering with games that they had



(a) Correlation: SportsAnno(0.577045482) Annoby(0.597640313)



(b) Correlation: SportsAnno(0.53815) Annoby(0.949586018)

Figure 43: The number of new threads started per user vs. replying to others' threads (a) with and (b) without Ireland and England Vs. Paraguay

not watched live. Only 25% said they used the system primarily to watch highlights of games they had missed live. Far more important was the commenting aspect of the system with over 85% of users saying the primary reason for using the system was to share comments about the games they had watched live on TV, closely followed by browsing the opinions of other users. The viewing of highlights clearly played an important part however in reminding users of important event. We can see from Figure 44 that keyframe clicks far outweigh the amount of annotations made per game. This is as a result of users who browse but do not annotate (∼50% of active users), as well as those who made annotations after viewing the associated highlights.

The fact that users of the system preferred to comment on the games which they had seen live was interesting since only 1 of the surveyed user had seen more than half the games broadcast during the competition. Most had seen between 6-10 games. Usage

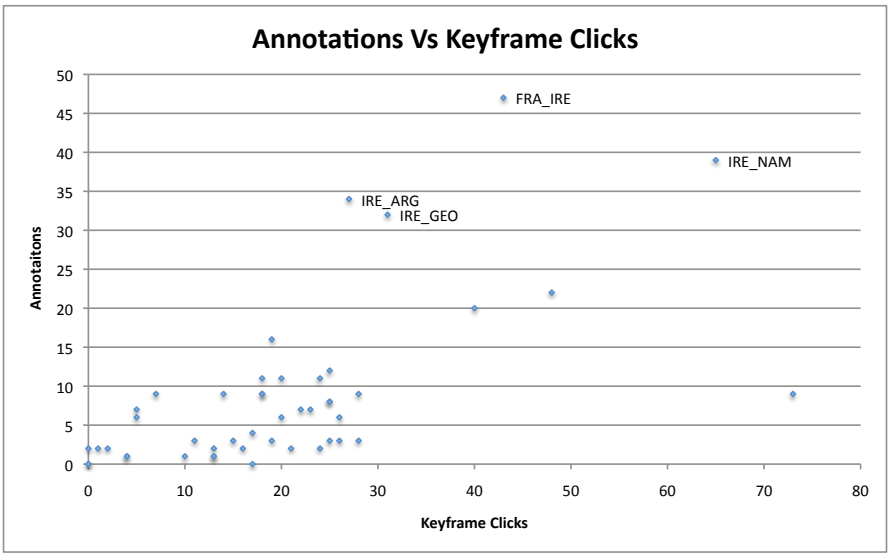


Figure 44: The Number of Annotations Made Vs the Number of Keyframe Clicks Per Game

of the system was typically between 5-15 minutes, with the majority of users stating they had used the system as a way to view highlights. This short usage time is seen as beneficial to the ideas used when creating the system. Both Annoby and SportsAnno were designed to facilitate summarisation of events within the original media, and community participation. The short usage times are seen as an indication that users did not see the need to spend large amounts of time browsing through games to find what they needed. Perhaps if some form of instant messaging service similar to those of social networking sites[16] was integrated into the interface, users may have spent longer on the system. The lack of instant feedback meant that users were more prone to checking the system for new information rather than browsing for extended periods.

All users surveyed said they followed sports regularly, watching highlights or live broadcasts on the television. Afterwards, users went to internet forums and websites to find more in-depth analysis and commentary. It was frequently stated that the ability to annotate and view matches in the same place was of great benefit.

Requested additional features for Annoby were similar to those which were requested for SportsAnno and not implemented. The creation of threads which were not anchored to any specific point within a report, allowing for creation of general conversation was desirable. Another important requirement is notification of replies (i.e. e-mail updates) to a user's comments. As mentioned with SportsAnno and previously (Brush et al., 2002; Cadiz et al., 2000; Sannomiya et al., 2000) the lack of a notification system can reduce the amount of interaction undertaken with the system by users. It also requires additional effort on the part of the users to re-find their own comments and check for

[16] http://blog.facebook.com/blog.php?post=12811122130

replies. Analogous to this, the ability to see all comments made by a specific user throughout the system was also requested, allowing for tracking of particular users by others and the creation of a more formal social network.

Resizing of the video playback window was also requested, along with a time-line which better indicated where commenting was taking place temporally within the context of each game.

3.3 Conclusions

The two systems we have built have enabled us to explore the requirements and attractions of a Web 2.0 annotation system. Through the creation of SportsAnno and subsequent refinement of ideas and presentation within Annoby, we have learned a lot about what features are most necessary to allow community interaction and participation. These systems have not however enabled us to built an annotation corpus large enough to fully and robustly test the algorithms which are the focus of this thesis. In the next chapter we shall explore an alternative approach to annotation corpus creation, a more simulated and synthetic approach. SportsAnno and Annoby have allowed us to attempt the creation of a real-world and natural annotation corpus created by real users. The inability to attract a large enough user community (due to constraints beyond our control such as copyright) means that we are unable to test the algorithms presented in Chapter 4 on this corpus except as a proof-of-concept. In order to test the scalability and robustness of these algorithms, we have been forced to turn to the approaches described in the following chapter.

CHAPTER IV

EXPERIMENTAL SETUP

The user experiments conducted using the SportsAnno and Annoby systems did not provide the volume of comments necessary for a thorough investigation of the hypotheses proposed in this thesis, though they did hint at the usefulness of both truly multi-dimensional browsing systems, and user-generated content in improving users' browsing experiences. As an alternative we have used the citation network of the SIGIR conference archive as a corpus. This substitute exhibits many of the characteristics of the smaller SportsAnno corpora, leading us to believe that it is indeed suitable as a corpus representing user's annotation and comments. We present comparisons and justifications for this in Section 4.2, comparing the reasoning for citation with that of annotation. Being a corpus made up of highly regarded and high quality scientific publications the calibre of the corpus is, like the newspaper articles annotated with the aforementioned systems, high in terms of quality of text. In place of the comments created by users in SportsAnno and Annoby on newspaper articles, we have created pseudo-annotations based on the citation of articles from within the scientific community as a whole to articles in our SIGIR corpus. The citing of papers by other authors has been shown to fit well within the general definition of annotation, and more specifically the annotation of SIGIR papers (Agosti et al., 2007). The source of citations to SIGIR papers is not as regulated as the original SIGIR publications since there is no guarantee that citations will come from as highly-regarded and strongly peer-reviewed conferences. This again echoes the trust-divide within the SportsAnno corpora between newspaper articles and users' comments. This SIGIR data provides us with the environment in which we can

develop and test our techniques. In this chapter we present a background to citation analysis as well as analysis of the SIGIR proceedings and its past uses. Lastly we present the algorithms which we have developed and which will be used to mine the data for information, and revisit the main hypotheses of this thesis.

4.1 Citation

When authors provide citations to previous work in the scientific domain, it is in an attempt to ground their new work in the context of past publications (Ziman, 1968). By citing previous work the author is explicitly conveying to readers the connection between the current work and that which has come before it. The way in which this connection is created however, can differ dramatically. Garfield observed 15 different reasons for the citation of articles within a work, ranging from paying homage to pioneers; giving credit for related work (homage to peers); and substantiating claims, to disclaiming or disputing previous work; criticising previous work; or correcting one's own work (Garfield, 1965). This classification however has been seen by some to over-simplify the reasons for citation. Brooks notes that the reasons for citation may in fact be far more complex than this, providing 7 categories[1] into which citation justification may fall (Brooks, 1986). The 7 justifications act as pieces of a jigsaw, combining in various ways to provide a more complex understanding of an author's reasoning in citing particular work. Brooks shows that authors will often cite a previous work for many reasons simultaneously, providing a contextual justification for the citation. When citing in a negative fashion, some authors have been shown to attempt to ameliorate the negative reference by providing positive support at the same time.

Citation, as with annotation, provides an author with an opportunity to create new information related to the original document. This new information can be of a critical nature or may in fact be neutral, providing further explanation or analysis of the idea put forward in the cited document. Citation context provides additional information about the reasons for citation which may aid in the information retrieval process, allowing the

[1] Brooks' collected justifications for citation:

1. Currency: Referencing the most up-to-date publications in a research field
2. Negative Credit: Criticism and other forms of debate of previous work
3. Operational Information: When others' algorithms, systems etc. are used within the current research
4. Persuasiveness: Grounding one's own work in similar research so as to persuade other of its validity
5. Positive Credit: Homage, validation or credit given to past works
6. Reader's Alert: References to background reading and leads to further research possibilities
7. Social Consensus: Referencing for a sometimes vague notion of consensus amongst peers

user to see more precisely why a citation was made. Without this context, the reader is left to guess why an explicit link has been created between two documents. Another advantage of citation context is that it is by its nature focussed on the material of the cited document and non-ambiguous in its relationship with that document.

4.1.1 Citation Indexing

Citation indexing is the process of indexing all citations made by articles, providing a means of discovering the relationships between articles and of inferring importance such as impact. The advantage of citation indices is that they allow for the identification and discovery of publications by topic, citation count etc. and not just through keywords, title, author etc. The indices also aid in navigation between papers through citation links both forward in time (moving from paper to referencing paper) but also backwards (through paper to cited paper) making indexes invaluable when performing searches for publications.

There are only a few citation indexes publicly available and the majority of these are commercial. The Institute for Scientific Information (ISI) provides a number of different scientific indices. Created by Eugene Garfield in 1960, the original ISI index was the first of its kind and has spawned several others. One of its indexes, the Web of Knowledge index[2], contains information on over 13,000 journals and 192,000 conferences ranging from science, social sciences and the arts. Other commercial indexes include the Westlaw index of legal publications. Google and Microsoft now have their own citation indexes which are freely available to web users[3]. All of these citation indices aim to return to the user a list of publications relevant to a query for author names, title, research aim etc. This is achieved through standard text-search but also incorporates linkage analysis (see Chapter 2). One additional advantage of the index however is for browsing. A user may browse the index not only for particular authors, but also by following the citation and reference links within the collection. The approach allows for a more serendipitous discovery of articles whilst removing the burden of providing highly-specific or overly-general queries from the user.

Citation indices were initially used in the context of information retrieval (Garfield, 1997). Within the context of this thesis, a citation index has been created which allows for the discovery of papers in response to a user query. The index itself however is hidden from the user and the information it provides is instead used to aid in the ranking of relevant documents/articles for the user query. The links between articles (created through citation and referencing) create a graph analogous to that of a collection of web-pages. The linkage graph is a directed graph (See Section 4.4) with each directed-edge

[2] http://isiwebofknowledge.com/currentuser_wokhome/cu_aboutwok/

[3] Google Scholar is still highly used, Microsoft's Live Academic is no longer available.

representing the citation of one article by another, and the vertices being the articles themselves. By looking at the graph of citations it is possible to discover article impact as well as the currency of an article in the recency of citing articles.

As well as providing information on specific articles, the graph may be used to discover information and statistics on the authors of each article. This additional source of information is of importance when it comes to returning a ranking for a user's information need. Past work on citation analysis has focussed mainly on the linkage structure of citation and has ignored the actual context of each citation. In terms of deciding which article is of more importance or relevance to a query, one easy measure to use is that of citation count. A paper which is highly cited within a field of research may be though of as being of greater merit and value than one which is cited less often (Peritz, 1992), provided issues regarding citation importance are taken into account. Even the position of the citation within a document as a measure of its intrinsic importance has been discussed (Cano, 1989; McCain and Turner, 1989). The manner in which each approach measured distance however differed greatly with the former measuring the percentage distance of the citation within the article, and the latter observing the section in which the citation was found.

Citation frequency was the earliest form of citation analysis. Many of the problems with performing a ranking based simply on citation frequency or count are covered by MacRoberts and MacRoberts (1989). One other obvious issue with citation frequency occurs when two articles are cited an equal number of times. When citing work within an area, authors create a form of preferential attachment (Jeong et al., 2003). This fact tells us that researchers prefer to cite those papers which are already well cited within the community.

This preferential attachment is reflective of the author of the paper also, leading us to believe that when performing a ranking, users would prefer to see equally cited papers ordered by author influence or presence. By this we mean that results would be ranked partially in terms of the overall impact which the author of a paper has had. For this to happen, statistics on the authors of papers must also be maintained containing information on the number of articles by an author within the index, a citation count of all the author's publications, recency of citations etc. With this information it is possible to differentiate not only the most influential authors within the index, but also the articles written by these authors which have made the greatest impact.

In this chapter we look at the act of citation from a direction comparable to the scenario proposed by the SportAnno project. That is, one which looks towards previous work (in the context of SportsAnno this implies previous comments) as a springboard for further inspiration or debate. These ideas fit well within both the categorisation provided by Garfield (1965) and also the refinement of Brooks (1986). While we intend

to focus on the linkage provided by citation, we shall also take into account the context in which citations are made. We do not however make any attempt to discover the exact reasoning behind a citation.

4.1.2 Citation Analysis

While it must be remembered that we have chosen to use citation and authorship of scientific papers as a means of creating a pseudo-annotations corpus, the area of bibliometrics and citation analysis has received much attention in the past. Its origins have been discussed in Section 4.1, but we feel that it is appropriate to also discuss some of the newer, popular measures from the field. These measures have been used to provide weights and authority to authors of papers through the observation of the affect an author's publications have had on their research field. Garfield originally proposed the use of citation counts and publication output as simple but effect measures of a researcher's prominence within a field. These measures however miss many of the subtleties associated with the citation and referencing of scientific papers. As a result, recent years have seen a number of suggested measures proposed, each of which attempts to capture more of the information about each citation, rather than just the number of citations. The most widely accepted of these measures is Hirsch's *h-index*, as well as the *g-index* which is a direct variant of the former.

4.1.2.1 The h-Index and Variants

Hirsch originally introduced the h-index (Hirsch, 2005) as a way to measure the impact of researchers within the physics community; he had noticed that factors such as number of papers and number of citations per paper did not sufficiently distinguish researchers who, say, had won a Nobel prize. Hirsch noted several problems and disadvantages to the approaches being used to differentiate the calibre of physicists. Some of these suggested measures where:

- **Number of Publications:** This will give a good indication of the contributions that a researcher has made, but provides no measure of the impact of any of these publications. In the context of our extended SIGIR corpus, this equates to the number of annotations/citations made
- **Total Number of Citations:** Whilst this will provide the missing information on the impact a researcher has had, it does not take into account the number of co-authors. A researcher who has co-authored with many different people can expect to have a higher number of citations. The effect of self-citation in this way was studied by Schreiber (2007). In our work we do not consider the affect of the

number of authors on a paper; in a real web annotation context, there would only be one author of an annotation. Self-citation has however been shown to more strongly affect the h-index of newer, less highly cited authors.

- **Number of Citations Per Publication:** This allows for a more even comparison of researchers of a different age, since the number of publications is not taken into account. It does however affect the calculation, as this measure favours researchers who produce less papers, but which have been more highly cited. In this way it is a bad measure as it discourages publication for fear of reducing the average number of citations per publication.

- **Number of Significant Papers:** *The number of papers with more than y citations* does counter all of the disadvantages of the above approaches. The arbitrariness of y however, means that different researchers can be greatly effected by different choices of y. It may also be necessary to introduce different y for different levels of seniority. (On reason for this is that in many areas of research, senior researchers may have their name attached to several papers which they have not directly contributed to.)

- **Number of Citations to Most Significant Papers:** *The number of citations to the q most-cited papers* counters the issues introduced above. Again, a new issue is introduced analogous to that of the last measure: q is arbitrary and may effect different researchers in different ways.

Hirsch's h-index attempts to counter the problems introduced by each of the above measures by setting a definite total on the number of papers to be considered when judging an authors contributions. It considers only those papers that are significant enough to have received a predefined number of citations. It is defined as follows:

Definition 2 *A scientist has an **h-index** of h if h of his total publications have received at least h citations each.*

By taking into account the number of citations to the top-cited papers, the h-index enables us to see the impact of a researcher on their field of research. The criterion that h papers be used however, removes the arbitrary qualities of the above variables y and q. This allows for a more realistic comparison of authors of different ages. The requirement that all h papers have at least h citations also means that an author who publishes a single work which is very highly cited does not receive extraordinary credit for this single work. Since it only takes into account those papers which are within the *Hirsch core* (Rousseau, 2006), it is not effected by the number of uncited papers, nor papers which are very highly cited.

While the h-index has become widely used in academic circles for the comparison of academics in search of tenure etc., it has been shown to have many flaws which make it a less than ideal comparison metric. Sanderson (2008) looked at the h-index of academics within the United Kingdom, using statistics gathered from different sources. This study highlighted the unsurprising affect of what publications/sources are included when measuring the h-index of a researcher. As with the most basic citation measures, it also suffers from being time-dependant: a researcher's h-index is strongly affected by the amount of time he/she has been in research (Burrell, 2007). Hirsch himself proposed dividing the h-index of an author by the number of years they have been publishing research to enable a fairer comparison of researchers of varying age. The field in which the researcher works can strongly affect the h-index; some fields such as physics can have hundreds of authors on a single paper. This fact when coupled with the effects of self-citations can lead to massively inflated h-indices for certain researchers. Much work has been done of scaling of the h-index across different fields (Batista et al., 2006; Iglesias and Pecharromán, 2007), taking into account both the average number of authors within the field and also the average number of citations. Iglesias and Pecharromán (2007) give the striking example of Andrew Wiles within the field of mathematics; Wiles' highly-celebrated proof of Fermat's Last Theorem (Wiles, 1995) won him great praise and notoriety. Looking at the Thompson ISI index we see that his h-index is just 12 from 13 publications. The paper he wrote on Fermat however, is unusual since it is the work of a single researcher over 10 years and is 109 pages in length. This point alone illustrates Hirsch's own concern over using the h-index alone to measure a researcher's worthiness for something such as tenure.

The h-index itself has been extended in several ways; as well as scaling, topic-specific variants have been introduced. The *h-b index* is used to discover hot topics as well as the general interest within a field of research (Banks, 2006). In the context of our own work, we see one possible avenue of interest as the h-index of an author specific to a certain search query. This is similar to Bank's h-b index, were the number of years n and publications included within the calculation of h-b are limited by the returned result set. Other than changing the set on which the index is calculated however, there have been some important variants introduced. The *g-index* which gives additional credit to researchers who have been highly cited on less than h occasions, as well as the *m-index* which considers a slightly different aspect of an author's contributions are discussed below.

g-Index The fact that the h-index is invariant under the presence of highly cited papers which have already been included within the Hirsch core is considered a major flaw or drawback of the index. Once a paper has been accounted for within the h-index, the number of citations it accrues from future publications has no affect on the h-index itself. Egghe noted that the h-index puts a lower bound on the number of citations that

a researcher must have had to achieve a specified h-index; a researcher must have had at least h^2 citations to their work to have achieved an h-index of h. This however does not take into account that any one of these papers may have received far more than h citations. He notes that *"it is an advantage of the h-index not to take into account the tail papers (with low number of citations) but it should (being a measure of overall citation performance) take into account the citation evolution of the most cited papers!"* (Egghe, 2006). It was for this reason that the *g-index was introduced*.

Definition 3 *A scientist has a* ***g-index*** *g if g is the highest rank such that the top g papers have, together, at least* g^2 *citations. This also means that the top* $g+1$ *papers have less than* $(g+1)^2$ *citations.*

Case studies have shown that the g-index better measures the citation records of researchers than the h-index since it provides additional credit to authors who are not only well cited, but who have contributed seminal works to a field (Egghe, 2006; Schreiber, 2008). It is however similar to the h-index in that it is monotonically increasing; once a researcher has achieved a certain g- or h-index, they can never fall below this. A scientist who retires, or who becomes an inactive researcher will (at worst) retain the same index.

a-Index Taking into account the contributions of just the *significantly cited* papers is one way in which to measure the impact of the top publications of a researcher. The h-index seeks to identify the most productive core of an author's output in terms of most received citations and as such, defines a good bound on the significant papers. The *a-index* looks at average number of citations received by the papers within the Hirsch core of a researcher.

$$a = \frac{1}{h}\sum_{j=1}^{h} cit_j \tag{23}$$

where h is the h-index of the researcher, cit is the citation count of item j i the Hirsch core.

The a-index was first considered by Jin (2006), and subsequently adjusted by Bornmann et al. (2008). In their paper they note that the distribution of citations is often heavily skewed; for this reason they propose using the median and not arithmetic mean of citations within the Hirsch core as a measure. They refer to this quantity as the *m-index*.

Many different variants of the h-index have been put forward in the recent literature. For a more thorough overview of these variants, the reader is directed to Bornmann et al.

(2008). Here we see that the authors have discovered two main groupings of indices, each of which measures a slightly different aspect of a researcher's contributions. We have chosen the g- and h-indices which *"relate to the number of papers in the productive core"* of a researcher; how much does the author publish? In the next chapter we will also look at the m index of authors from our extended SIGIR corpus. The m index *"relates to the impact of the papers in a researcher's productive core"*; how many citations have these papers received? This is subtly different to the h-index since an author with just 3 papers cited at least 3 times will have an h-index of 3, but if these papers have been cited far more than 3 times the m-index will be far higher.

In identifying these two different groupings of h-index variants, Bornmann et al. proposes *"any pair of indices as a meaningful indicator for comparing scientists, where one index relates to the number of papers in a researcher's productive core and the other index relates to the impact of the papers in a researcher's productive core"*. The measures that we introduce in Section 4.6.1 consider both the contributions of an author to the discussions and threads within the extended SIGIR corpus, as well as the amount of interaction/citation which this contribution receives. In this way we feel that we have created a measure that is a meaningful indicator of an author's contribution.

4.2 An Annotated Corpus

The similarity between citation and annotation of documents has been noted in the past (Agosti et al., 2007). In citing a publication, the author is in some way acknowledging the role of past work in the current publication. This acknowledgement may be seen as a comment or annotation on the cited article. Annotations are used to, amongst other things, support the arguments within a document, or to expand upon and illustrate a point (Marshall, 1997). There are evident parallels between the motivations for publicly annotating a document (annotation which is designed to be read not just by the creator but also others) and citation practice. In fact, all of the aspects of citation practice may be considered equal to some aspect of annotation. Of the 7 reasons for citing a document presented by Brooks (Brooks, 1986) and mentioned in the previous section, the least comparable is social consensus. This however follows directly from the idea of threaded annotations or comments where the consensus is self-evident, as in Figure 45.

Due to the relatively small size of the annotation corpora produced during the SportsAnno experiments, it was necessary to find a larger corpus of annotations. A publicly available corpus of the desired type and size does not, to our knowledge, exist and so a suitable proxy was needed. While large-scale experiments have been carried out on collections of web-pages annotated by tagging (Schenkel et al., 2008; Halpin et al., 2007), this level of annotation or indeed any annotation which will involve just assigning tags to an object, does not provide the depth needed to use the techniques developed in this

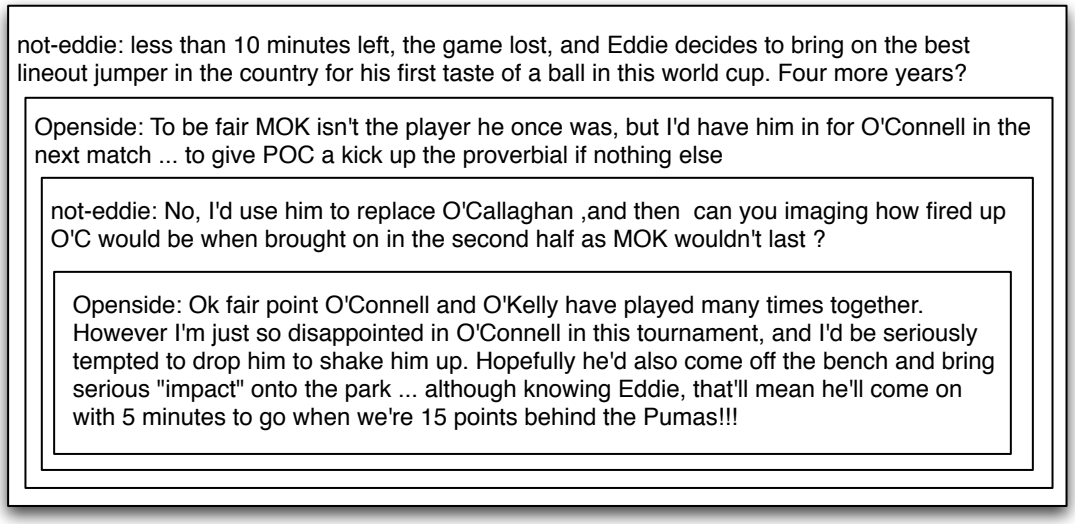


Figure 45: Threaded converstation from Annoby corpus

thesis. Tagging has been shown to provide a wealth of additional information which is helpful in satisfying a user's information needs, sometimes, but tags lack contextual information. Tagging does not provide the reasoning behind the tags applied, nor can the level of interest ascribed to the annotated document be gauged from these tags. These are two problems alleviated by the used of citation context and more, in the case of SportsAnno, the creations of phrasal-comments rather than single-word tags.

Citeseer (Bollacker et al., 1998) is a publicly available and well-known citation index originally created in 1997. Unlike other indices, it provides contextual information about the citation and referencing of over 1 million documents, constructing an index of papers along with citation context information for each citing document. Two facts to note are that this context is of a fixed size around the citation marker, and is obtained automatically. Unlike our index, if a paper can not be de-coded to provide a context, no context is provided (see Section 4.3.1). This index was created through a crawl of the web, seeded with pages returned from a search query to various web search engines for words like "publication", "papers" etc. (Giles et al., 1998). To be able to utilise a sub-collection of CiteSeer alone (limited to, say, a single conference proceedings) is not enough, since it requires that all the papers we wish to use are publicly available and found openly on the web which is not the case.

4.2.1 The SIGIR Corpus

As the world's largest educational and scientific computing society, the Association for Computing Machinery (ACM) provides the computing field's premier Digital Library.

Through the library, members have access to leading research publications and conference proceedings. Through the ACM portal, citation information has been gathered for the Special Interest Group on Information Retrieval (SIGIR)[4] conference on "Research and Development in Information Retrieval" series spanning the last 10 years, including publications which cite SIGIR articles, but which are not themselves from within the SIGIR proceedings.

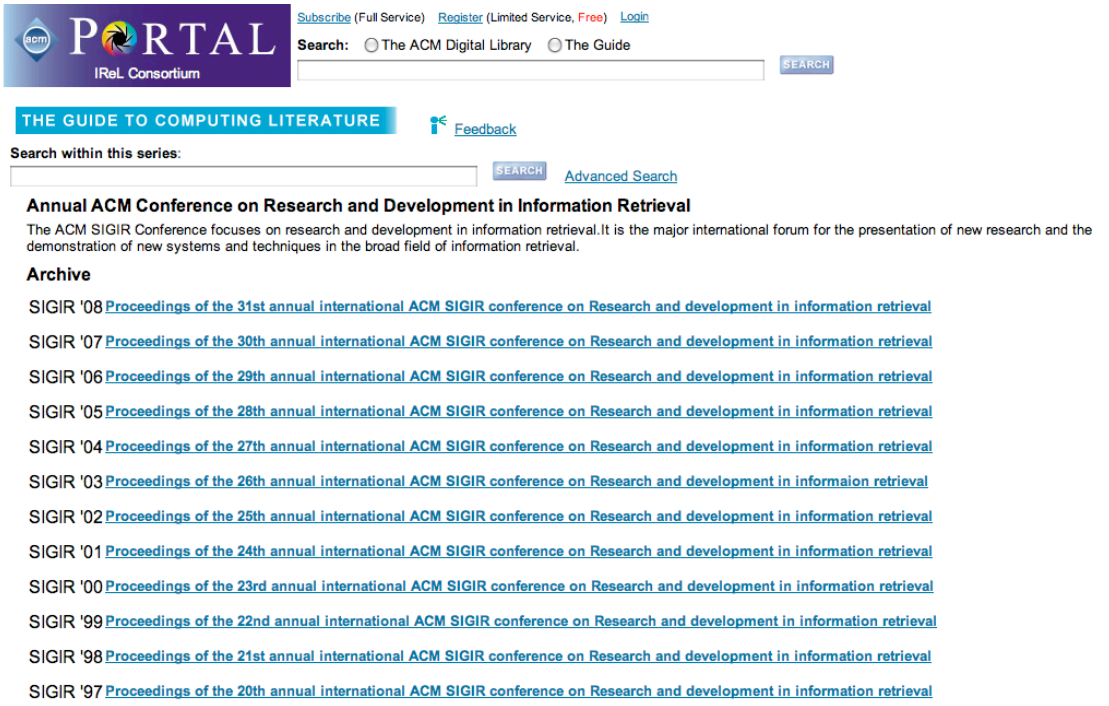


Figure 46: SIGIR archive in the ACM database

The method used to obtain and construct the annotation corpus which we use in our experiments, as well the building of a citation index, are discussed in 4.3.1. We believe that this corpus with its increased size alleviates the major drawback of the corpora created within SportsAnno; SportsAnno yielded too few annotations. We also believe that the inclusion of citing papers from outside of the proceeding themselves extends the work carried out previously on the SIGIR proceedings alone.

The availability of citation information is made possible, as previously mentioned, through the ACM database. This database holds all information regarding publication of articles (date, authors, citations, references). A digital copy of each of the publications

[4]The annual international Special Interest Group on Information Retrieval (SIGIR) conference series on Research and Development in IR (http://www.sigir.org/index.html), which began in 1978 (an initial SIGIR conference was held in 1971), is considered the most important in the field of information retrieval. The focus is on *"all aspects of information storage, retrieval and dissemination, including research strategies, output schemes and system evaluations."*. It is a highly selective conference with acceptance rate typically ∼20%, all papers having been peer-reviewed by several reviewers before acceptance. This low acceptance rate leads to papers of very high quality meaning the corpus as a whole is an excellent source of quality information and may be considered highly authoritative.

Table 5: PDF file error Statistics for the extended SIGIR corpus

	SIGIR	*Non-SIGIR*
Collection (N^o of Papers)	767	2115
Poster/Demo Paper	174	0
Corrupt/Unavailable Download	13	859
Missing From Reference Section	0	4
Missing Reference Marker	0	2
ACM Error (No Citation Made)	8	46
Papers Available For Analysis	572	1204

themselves is also available in PDF format for any publications from ACM affiliated conferences and journals which may themselves cite SIGIR articles. Where no publication is available, publisher information is provided. Table 5 shows the percentage of articles which cite SIGIR articles for which it was not possible to retrieve a digital copy of the publication, although in theory we could have searched the web as a whole, we chose to confine our corpus to those publications which are available directly through the ACM database. While this has led to an approximate 40% loss of citation contextual text for citations external to the SIGIR proceedings, it does not affect the citation information provided by internal SIGIR publications.

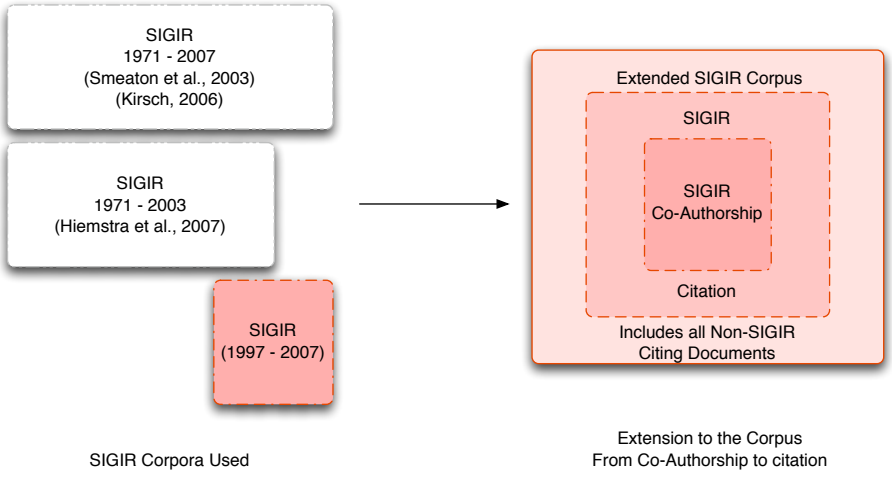


Figure 47: The scope of previous studies based on SIGIR proceedings. We can see that while the time-window which is used in our studies is smaller, the depth to which analysis is performed is far deeper

The SIGIR publications network itself has been used in research previously (Figure 47). Smeaton et al. (2003) were the first to perform a study on the database of SIGIR articles, using it to observe "hot topics" within the IR world as part of the 25th anniversary celebrations of the SIGIR conference, as well as to look at the co-authorship network within the SIGIR proceedings. Their work on the first 25 years of SIGIR was subsequently re-visited and extended by Hiemstra et al. (2007) who performed their analysis

for the 30th year of the conference's history, noting differences in the most-connected author within the network of SIGIR, and as well as the geographic prolificness of submissions. Kirsch (2006) used the PageRank algorithm (Page et al., 1998) to perform linkage analysis on the graph of SIGIR authors. The corpus has not however, to our knowledge, been extended to take into account the citation of SIGIR papers, by papers external or internal to SIGIR. This additional information allows us to build a more complete picture of the importance of individuals in the SIGIR publications network, and also provides a means of measuring the impact of any single publication (Salton, 1971a; Garfield, 1965). Again, CiteSeer does have the capacity to perform some of the corpus construction which has been done by the authors, though only for those papers which are published openly on the web.

By extending the SIGIR corpus to include the citing articles, the corpus becomes a far better approximation of the corpora created by SportsAnno and Annoby. By extension, we have also created a corpus that is a good approximation of general Web 2.0 tagging, commenting and annotation as well of blog commenting and track-backs. These interactions are represented here as citations upon original documents, much the same as tagging of pages or commenting on blogs. We have chosen SIGIR in particular due to the previous work which has been done on the corpus (allowing us to perform some comparative studies) and also due to its highly cited nature. The SIGIR Conference ranks within the top 6% of Information and Communications Technology (ICT) conferences according to a study performed annually by the Computing Research and Education Association of Australasia (CORE)[5]. As a comparison we have measured the average number of citations received by both SIGIR and non-SIGIR articles within our corpus. SIGIR articles receive a mean average of 3.39 citations per publication. Non-SIGIR publications within our corpus receive just 0.29 citations per article, from all sources, assuming non-SIGIR papers citing SIGIR papers are representative of the average paper in Computer Science. SIGIR's strong citation characteristic provides a wealth of citation context to use in the experiments described in Chapter 6.

We have constructed our corpus from a 10 year window of the SIGIR proceedings, ranging from 1997-2007. Within this collection there are over 4000 authors, ∼770 SIGIR publications and an additional ∼2100 non-SIGIR publications which cite these SIGIR articles. The publications which are not from within the SIGIR proceeding come from the ACM database of publications[6]. Within this database are all publications from conferences affiliated with the ACM, consisting of many of the top ranking journals and proceedings of the Information Retrieval domain.

We originally attempted to obtain the full citation information for all 30 years of SIGIR, but we were forced to reduce our collection to that of 1997-2007. The main

[5] http://www.core.edu.au/
[6] http://portal.acm.org/portal.cfm

reason for this was the difficultly in retrieving correct citation context information for papers published earlier than 1997. The volume of manual cleaning of the data required was prohibitively high so the collection-size was fixed at 10 years. We also choose to disregard SIGIR poster and demonstration papers. These papers are not as highly cited or reviewed and in general describe work which is less in-depth or complete.

4.3 Deriving Annotations from Scientific Citations

During an extensive analysis of the properties and uses of annotations, Agosti et al. (2007) mentions several properties of annotations which are analogous to the reasons for citation. As meta-data citations perform the same task as annotations, providing additional information on the original document. They may also be seen as a hypertext connecting citations together; when several citations are made concurrently within a single sentence or context, these cited documents are implicitly connected in some way. Finally, citations can provide an additional layer of context as annotations may do; they "can make hidden facets of the annotated documents more explicit" or clarify and refute a conclusion.

The citing of documents is done for a number of reasons, but without any form of context for the citation it is not possible to understand why a citation has been made. Citation context provides the extra knowledge required to decide upon the merit of the cited document, as mentioned in Section 4.2. As with the Citeseer project, we have taken the context of a citation into account as well as the linkage structure between cited and citing documents in our work. In order to discover this context, we performed a number of steps on the texts of both SIGIR and non-SIGIR documents which cite SIGIR documents. Figure 48 gives an overview of the system developed to retrieve and clean documents for analysis.

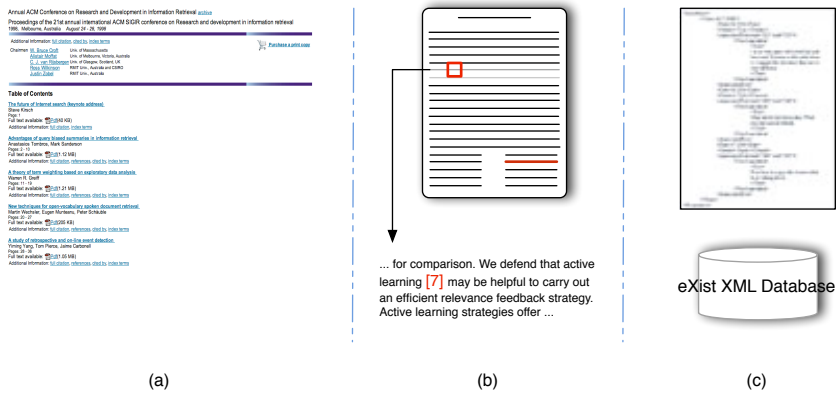
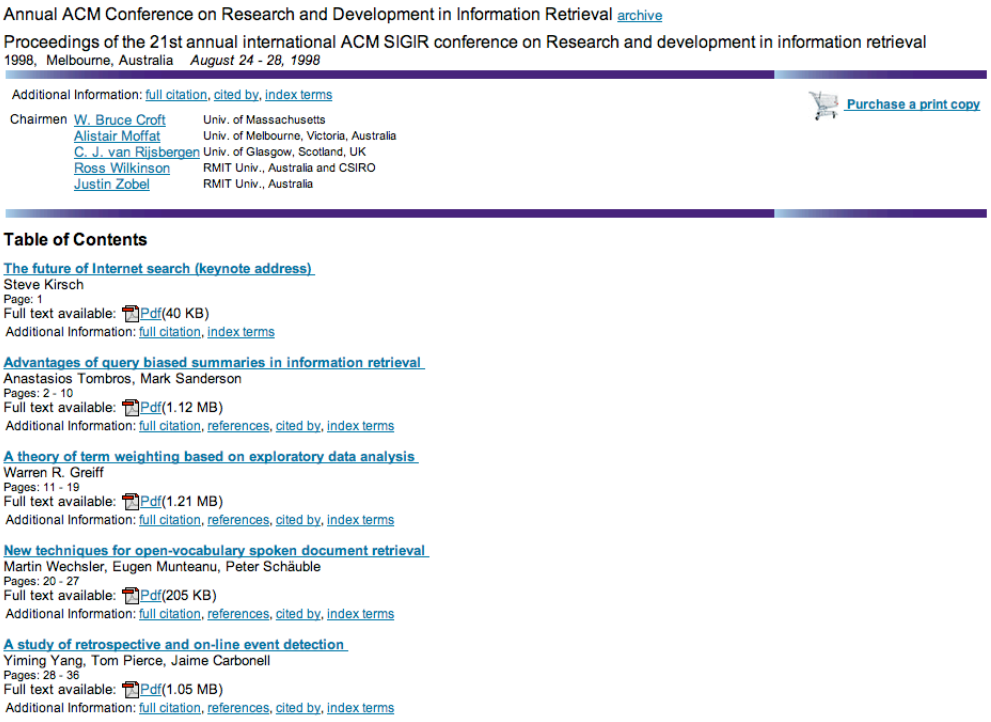


Figure 48: Creation of the SIGIR corpus: (a) Download (b) Citation Recovery (c) Storage as XML

4.3.1 System Description

Download Starting at the SIGIR main proceedings archive within the ACM portal website (Figure 46), links are followed to each SIGIR conference year. From here we have a list of all the papers within the conference each year, with links to each paper. The web-page for each SIGIR paper contains all the information and links to referenced and citing papers. By following these citing paper links we arrive at the page corresponding to that paper. This is how the tree of citations is traversed.

It is possible for a loop of citation to exist where-by a paper is cited by a paper that it itself cites, or one of it's citing publications cite. This problem is dealt with here at the download stage by simply following links from the initial publications page only. At this point, we only download the publications which cite the paper we are currently interested in. In order to create the threads of citations we are interested in, we later check for links between citing publications of different papers. This is explained in section 4.3.2.


Annual ACM Conference on Research and Development in Information Retrieval archive

Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval
1998, Melbourne, Australia *August 24 - 28, 1998*

Additional Information: full citation, cited by, index terms

Purchase a print copy

Chairmen W. Bruce Croft — Univ. of Massachusetts
Alistair Moffat — Univ. of Melbourne, Victoria, Australia
C. J. van Rijsbergen — Univ. of Glasgow, Scotland, UK
Ross Wilkinson — RMIT Univ., Australia and CSIRO
Justin Zobel — RMIT Univ., Australia

Table of Contents

The future of Internet search (keynote address)
Steve Kirsch
Page: 1
Full text available: Pdf(40 KB)
Additional Information: full citation, index terms

Advantages of query biased summaries in information retrieval
Anastasios Tombros, Mark Sanderson
Pages: 2 - 10
Full text available: Pdf(1.12 MB)
Additional Information: full citation, references, cited by, index terms

A theory of term weighting based on exploratory data analysis
Warren R. Greiff
Pages: 11 - 19
Full text available: Pdf(1.21 MB)
Additional Information: full citation, references, cited by, index terms

New techniques for open-vocabulary spoken document retrieval
Martin Wechsler, Eugen Munteanu, Peter Schäuble
Pages: 20 - 27
Full text available: Pdf(205 KB)
Additional Information: full citation, references, cited by, index terms

A study of retrospective and on-line event detection
Yiming Yang, Tom Pierce, Jaime Carbonell
Pages: 28 - 36
Full text available: Pdf(1.05 MB)
Additional Information: full citation, references, cited by, index terms

Figure 49: Conference main page listing all the papers and providing links to each paper

On the citing papers page, we check for a link to the full-text PDF copy of the paper. If this copy is available, the paper is downloaded and processed. If no PDF is available, we still make note of the number of citations made to this paper. This fact is used in the calculation of a publications value. We do this for all citing papers, but for non-SIGIR papers this is the only information which is stored regarding citation. (No specifics on what papers link to the current paper are stored for non-SIGIR publications.)

In order to limit the time taken to perform the downloading of papers, as well as the size of the downloaded collection, we limit the number of non-SIGIR papers which are downloaded per paper to a maximum of 20. All SIGIR papers themselves have already been made available from Smeaton et al. (2003) and so there is no need to download the PDFs of these papers. For any given paper, all PDFs of citing SIGIR papers are available, as well as a maximum of 20 non-SIGIR PDFs. The links to all other citing documents are however stored for future exploitation. We also take the full number of citations (if greater than 20) into account when calculating ranking of a paper later on ensuring that highly cited non-SIGIR papers are still given appropriate credit or weight.

PDF decoding The downloaded PDFs must first be converted into text so as to perform text processing. This decoding was done using the *PDFbox*[7] toolkit. Machine readable PDF has not always been the norm and so we were unable to decode some documents since the PDFs consisted of scanned images of pages from the proceedings. This was the case for the majority of papers within the SIGIR proceedings prior to 1980, and also affected a number of papers external to the SIGIR proceedings. Table 6 shows the number of papers where the PDF could not be de-coded. Regardless of whether or not the paper is decoded, a record is created for each publication.

Table 6: PDF file error Statistics for the extended SIGIR corpus

	SIGIR	*Non-SIGIR*
Collection (N^o of Papers)	767	2115
Poster/Demo Paper	174	0
Corrupt/Unavailable Download[8]	13	859
Missing From Reference Section	0	4
Missing Reference Marker	0	2
ACM Error (No Citation Made)	8	46
Papers Available For Analysis	572	1204

Citation Recovery To complete our citation network we need to recover citation links by processing the PDF text. We do not face the same problems encountered by Giles et al. (1998) regarding identification of citations to the same paper. In our collection, the title of the each paper is contained in the anchor-text for that paper's web-page. The way in which a citation is presented within different papers (both SIGIR and non-SIGIR) however differs greatly. Figure 50 shows just some of the ways in which the same paper is cited within the collection, despite there being an ACM "house-style" and clear guide-lines for authors. The vast majority of papers are cited in one of the first two ways, but in order to obtain as much information as possible, all citation-methods are catered for:

[7] http://www.pdfbox.org/

[8] Nick Craswell , David Hawking , Stephen Robertson, Effective site finding using link anchor information, Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, p.250-257, September 2001, New Orleans, Louisiana, United States

Craswell N., Hawking D., Robertson S. (2001), Effective site finding using link anchor information, Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, United States

8. N. Craswell , D. Hawking & S. Robertson, Effective site finding using link anchor information, Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, p.250-257, September 2001, New Orleans, Louisiana, United States

[CRA01] N. Craswell et al., Effective site finding using link anchor information, Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, p.250-257, September 2001, New Orleans, Louisiana, United States

Figure 50: Referencing styles

1. **Title Discovery:** The reference to the paper of interest will always contain the title of the paper being referenced and at least the first author's surname or family name. As shown in Figure 50, the names of other authors on the author list may be abbreviated. We therefore identify firstly the correct reference from within the reference section. Once this is done, the reference marker itself must be found. This is the means of identifying the reference from within the main body of text of the paper. We have developed a number of java regular expressions which identify first the title and then the reference marker.

2. **Reference Location:** References are located within the main body of text using the reference marker retrieved from the previous step. Normally, the reference marker retrieved from the reference section is used within the main body but on occasion a mixture is made between reference section and main paper. By this we mean numbers may be used in the reference section but author names in the main article and vise-versa. For this reason, we first search for the title of the cited paper within the reference section and then make note of both the cited authors' name and the reference marker used. If a search for the marker is unsuccessful, a follow-up search is performed for variations and abbreviations of the author-list.

3. **Context Retrieval:** The window around the reference marker which should be taken as context for the citation is chosen heuristically. The sentence in which the

reference is found plus the sentences before and after, are used to create the context of the reference. Other more sophisticated algorithms such as the text-tiling algorithm (Hearst, 1997) may return better results, but as a heuristic this method appears to work adequately. Recent work by Ritchie et al. (2008) shows that this choice of context-window is good for discovery of index terms for the cited document. This fact leads us to believe that it is also a good choice for citation-context length. The length of the context itself is of interest to the authors in simulating the random-length commentary which took place during the SportsAnno experiments. This feature of the system (the choice of fixed sentence-length as opposed to fixed character-length context) is examined in more detail in Chapter 6. Note that on occasion reference markers may appear within tables. If this is the case then the citation context is still chosen in the same way leading to slightly large contexts which take the whole table into account. This does not happen often and so is not considered an issue.

4. **Recording:** Once the context of the reference has been retrieved, the information on this citing document must be added to the record of the SIGIR document it has cited. In this file all information about the associated document is stored. This includes:

    - *Title and Authors:* The name of each author is stored, along with the position of the author within the author list. (Author names are sanitised so as to remove ambiguities from the final collection i.e. W.B. Croft and W. Bruce Croft)
    - *Citation count:* This is the full citation count of the paper, including all non-SIGIR citations which were not downloaded as a result of reaching the 20 paper maximum.
    - *URL:* This is the URL of the paper. The URLs of all the non-SIGIR papers which were not downloaded are also recorded so as to provide an opportunity for future expansion.
    - *Citation:* This is the information about the citing paper, including all the information already mentioned (title, authors, URL etc.) as well as the citation context. If a paper is cited several times within the single paper, multiple context nodes are created, one for each citation context.

Storage: The above process is repeated for every citing paper creating a complete XML file (see Section 4.3.2) for each SIGIR paper which contains all the document information, as well as the citation information required to construct a citation graph. This is discussed in Section 4.4. A text file is also created as a by-product of the PDF-decoding process. This file is also stored. In this way we have created in effect, two related collections. The collection of text documents contains all the text from the papers

within the collection, enabling full-text information retrieval against the collection. The second collection contains all the citation information about the collection, including linkage structure and citation contexts. This second collection allows for the construction of a graph in which papers form the nodes, and citations from one paper to another form edges. These edges may then be weighted according to a combination of factors retrieved from within the XML files.

4.3.2 An XML Collection

An XML document is initially created for each SIGIR document within the collection. Each paper's XML file is disjoint from all other files; every time the paper is cited, the citation-context from the citing paper is inserted into the XML file as a 'Citation' node. While the non-SIGIR citations are finished with, the citations to SIGIR papers by other SIGIR papers may be used to create a graph. By replacing the SIGIR 'Citation' node within an XML file with the citing SIGIR paper's entire XML file, we are able to create a thread of citations which is ordered in ascending chronological order. This effect is known as "threading" and is regularly found on internet forums and within blogs. This idea is illustrated in Figure 51.

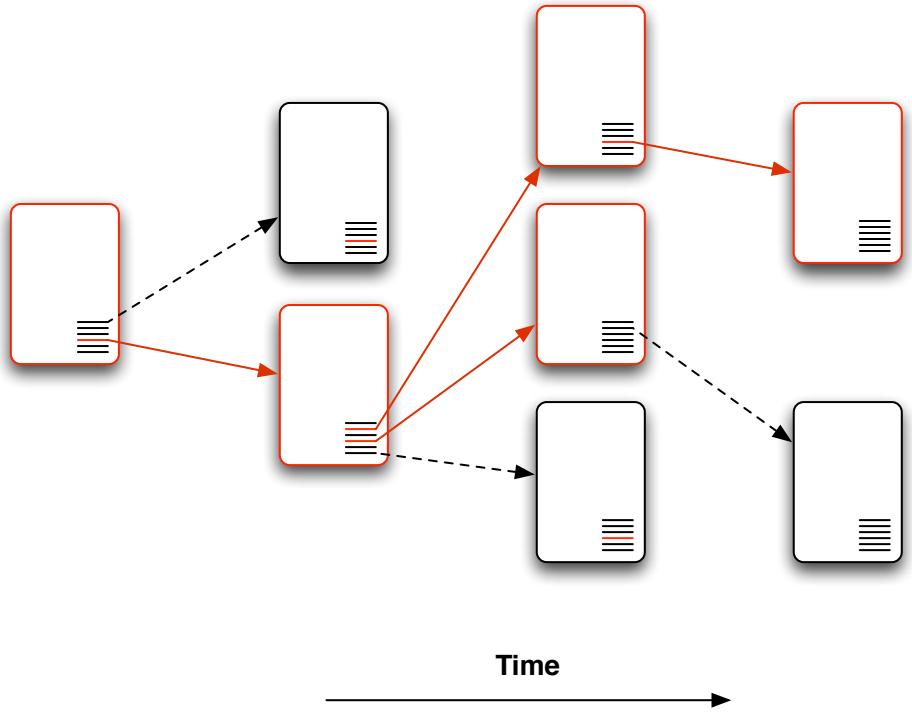


Figure 51: Linking of documents via citation. Each SIGIR (red) document's links are followed. The dotted links are to non-SIGIR documents. In this example, only some citation links are created; in reality all links are created.

Once threading is achieved, papers may be considered in the context of their thread;

papers which cite a paper are parents of the paper; papers which are cited by a paper are children of the paper.The entire thread above the paper of interest is the ancestry of the paper. This process of threading the SIGIR papers into the eldest ancestor's document reduces the corpus of documents from the initial ∼570 down to 251 unique documents. This is a result of SIGIR papers citing previous SIGIR papers. If a SIGIR paper has cited a previous SIGIR publication, and as a result has had its XML file subsumed into the cited documents file, then the corresponding publication's XML file is removed from the collection.

One of the major benefits of XML storage is the structured nature of the documents which allows the querying of documents through both structure and content (Guo et al., 2003; Liu et al., 2004; Tatarinov et al., 2002; Wolff et al., 2000). By this we mean that the information has a structural value, with certain information becoming more important by virtue of its place within a document. This is important when discovering if the result of a text query, say, comes from the title of a document, or from the citations etc. We have not stored the text of each PDF in XML however, as the overhead to doing this was too great. It would be possible to extend the work presented in this thesis and in doing so enable structural queries against the SIGIR publications (e.g. only return results which are from within the abstract of a paper). This work is however beyond the scope of the current thesis, and also creates a disparity between the then highly ordered documents within our extended SIGIR corpus and the documents found, say, on the internet.

Another advantage of the technology is the straightforward extensibility of any document to incorporate new information. Using XML allowed us to easily add information which was calculated from the network as a whole after document construction. The independence of every document also ensures that there are no problems of relational ambiguity and duplication which can plague relational database implementations.

This threading of the citations allows for both the easy creation of a graph, and analysis of the evolution of a research idea. One important point that was made earlier was the reasoning behind not following the citation links when downloading papers is the discovery of link-sinks. With the re-creation of citation threads, this problem is re-introduced. In order to cope with it, any possible loop is prevented in the following way:

1. When combining two XML files together, first check if a citation node corresponding to the paper we are adding is already present within the ancestry of the node which we are adding to. If an ancestor is found, add the XML but do not attempt to follow any of the new children which have been introduced as a result of the added XML. This is to prevent infinite loops where citing documents are cited by cited documents etc.

2. If no similar ancestor is found, replace the Citation node within the file with the corresponding complete XML file.

3. Repeat step 1 for each of the newly introduced children.

Note that the combination of files must be performed in ascending chronological order from the earliest paper up to the most recent citation, and not by following references backwards. This is to ensure that papers which are referenced by other papers in the same year are not affected. If this is not done then these papers are not dealt with correctly.

All documents were stored within an eXist database running on a Windows Server 2003 PC with 4GB of RAM. The citation of earlier documents within the collection by many other documents did lead to the file-size of documents growing dramatically. While most documents were approx. 400-600kb, some files grew to over 10MB in size. The complex nesting within these documents required the memory allocation to be greatly increased for the database.

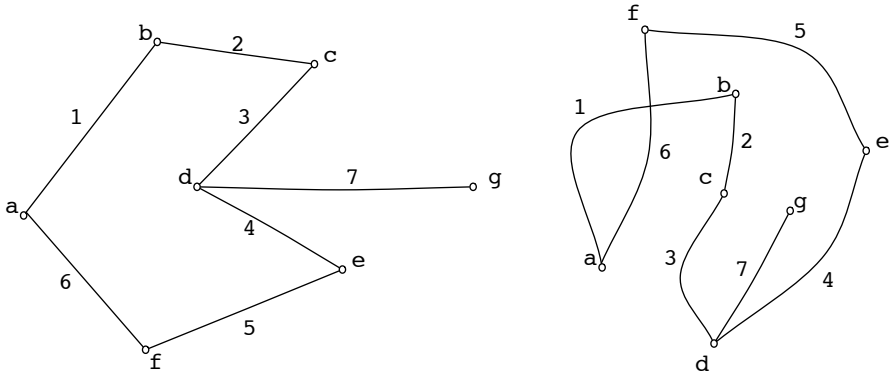
4.4 Graph Theory

We can represent many of the interactions that take place in both the physical and virtual or digital worlds by a series of distinct concepts joined together in some manner. These concepts can be anything from members of a company to food-types, with the manner of connection being acquaintance or ingredient for a meal. The manner in which these objects or concepts are joined is sometimes not really of any importance when we are more often interested in whether two objects are indeed connected. Drawn graphically, each of these objects may be seen as a dot, with connections being represented as lines or arcs joining them (Bondy and Murty, 1976).

Definition 4 *A **Graph** G is an ordered triple $(V(G), E(G), \psi_G)$ consisting of a non-empty set $V(G)$ of vertices, a set $E(G)$ disjoint from $V(G)$, of edges, and an incidence function ψ_G that associates with each edge of G an unordered pair of (not necessarily distinct) vertices of G. If e is an edge and u and v are vertices such that $\psi_G(e) = uv$, then e is said to join u and v; the vertices u and v are called the ends of e.*

An example of this may be seen in Figure **??**. Both are representation of the same graph with all relations remaining constant. The graph itself is an abstraction of the information it represents. Mathematically, it is easier to write the graph as follows:

$$G = (V(G), E(G), \psi_G)$$



(a) (b)

Figure 52: The graph of G drawn in two distinct ways

where

$$V(G) = \{a, b, c, d, e, f, g\}$$

$$E(G) = \{1, 2, 3, 4, 5, 6, 7\}$$

and ψ_G is defined by:

$$\psi_G(1) = ab, \psi_G(2) = bc, \psi_G(3) = cd, \psi_G(4) = de,$$

$$\psi_G(5) = ef, \psi_G(6) = fa, \psi_G(7) = dg,$$

A graph is **simple** if it contains no loops and no two of its edges join the same two vertices. As we shall see, neither the graph of author citations nor paper citations within our extended SIGIR corpus is simple. Authors frequently cite their own previous work which leads to loops within the graphs. Both graphs are however *finite*, since there exists only a finite number of vertices and edges within each graph.

Definition 5 *A graph H is a* ***subgraph*** *of G (written $H \subseteq G$) if V(H) $\subseteq$ V(G), E(H) $\subseteq$ E(G) and ψ_H is the restriction of ψ_G to E(H). If $H \subseteq G$ but $H \neq G$, then H is a proper subgraph of G. A spanning subgraph of G is a subgraph H, where V(H) = V(G).*

Graphs may be either directed or undirected. A **directed graph** D is an ordered triple $(V(D), A(D), \psi_D)$ consisting of a non-empty set $V(D)$ of vertices, a set $A(D)$, disjoint from $V(D)$, of *arcs*, and an *incident function* ψ_D that associates with each arc of D an ordered pair of (not necessarily distinct) vertices of D. If a is an arc of D, and u and v are vertices such that $\psi_D(a) = (u, v)$, the a joins u and v; u is the *tail* of a and v is the *head*.

The **degree** δ_v of a vertex v in G is the number of edges of G incident with v, each loop counting as two edges. When speaking of vertices in an *undirected graph*, we must

distinguish between the **in-degree** and **out-degree** of a vertex. The out-degree, δ^+, is the number of incident edges beginning at a node, while δ^- the in-degree is the number of edges terminating at a node:

$$\delta^-(v) = |\{[v', v'']\epsilon E \mid v'' = v\}|$$
$$\delta^+(v) = |\{[v', v'']\epsilon E \mid v' = v\}|$$

A **path** between two nodes v and v' is a sequence of nodes v_0, ... , v_k with $v_0 = v$ and $v_k = v'$ and $\{v_i, v_{i+1}\}\epsilon E$ and $v_i \neq v_{i+1} \forall i$. The *distance* between two nodes u and v is the shortest path between the two nodes. If we associate a real number $w(e)$ to each edge e of G, then G becomes a *weighted graph*. The **weight**, $w(e)$, of an edge may be taken into account in some functions which measure the distance between the tail and head of the edge. Weight is usually applied to a graph in order to signify an intensity of the connection between two vertices. In the case of the SIGIR-authors graph, increased weight may be used to signify co-authorship of a paper.

Two vertices u and v of G are **connected** if there exists a path between u and v (Godsil and Royle, 2001). A graph is connected if any two vertices can be joined by a path and the 2 variations of this are strongly connected and weakly connected, and are defined as follows.

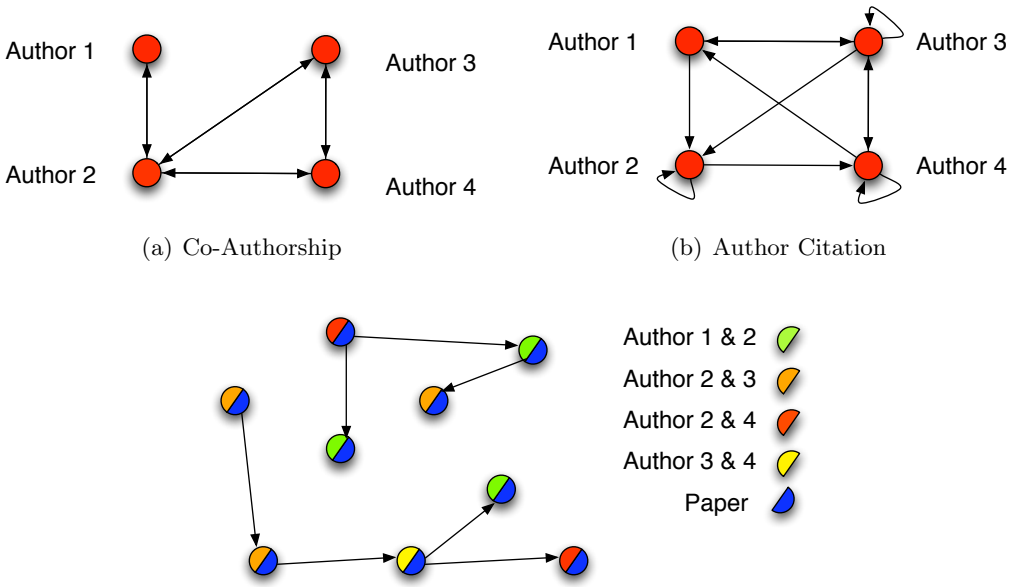
Definition 6 *A directed graph G is* ***strongly connected*** *if any two vertices can be joined by path of distinct nodes* $(u_i, ..., u_k)$ *such that* $u_i = u$ *and* $u_k = v$*. If only the underlying undirected graph is connected, then G is said to be weakly connected.*

4.5 Corpus Analysis

Our extended SIGIR corpus provides a collection of authors and publications on which to perform both citation and network analysis, as well as information retrieval. With 4000 authors and 3000 publications, it expands greatly on the corpora we obtained earlier through the SportsAnno experiments. In this section, the graph of the SIGIR proceedings as a whole is studied in detail, both in the context of author-to-author citation, and paper-to-paper citation. Also, measures of author co-citation are made so as to enable comparisons to previous work carried out on the SIGIR proceedings alone (Hiemstra et al., 2007; Kirsch, 2006; Smeaton et al., 2003).

Citation analysis involves the construction and analysis of detailed graphs linking the reference and citation of documents to one another. The citation and co-citation of scientific authors' papers has been used in the past to study network properties of the

scientific community, discovering information about research trends, social properties of the network and the influence of authors within the scientific community.



(a) Co-Authorship

(b) Author Citation

(c) Paper Citation

Figure 53: Example graphs showing the types of connections being made in the extended SIGIR corpus graphs

Table 7 shows statistics on the level of connectedness within our extended SIGIR collection. The 4 different graphs which are compared all consist of nodes and vertices from within the extended SIGIR corpus. The first two graphs have as vertices the authors of papers; in the first graph an edge is created between co-authors, similar to Figure 53(a); the second graph has edges between authors who have cited each other, as with Figure 53(b). The first graph is the co-authorship graph for the SIGIR proceedings alone. It spans the 10 years between 1997 - 2007 but includes only the authors who have published a SIGIR paper, excluding all authors of non-SIGIR papers within the extended SIGIR proceedings. In the later two graphs, vertices are papers and edges occur between a paper and another paper which it has cited, as with Figure 53(c). Only the first graph is undirected since the co-author relationship is bi-directional and reciprocal. The later two graphs are for the citation networks within the extended and original SIGIR networks respectfully.

Social Network Analysis (SNA) existed long before the advent of the internet, meaning many of the techniques used for analysis of the web as a whole find their roots in previous work within SNA. Link-based algorithms such as PageRank (Page et al., 1998) and HITS (Kleinberg, 1998) are very similar in their search for authoritative pages to bibliometric and citation analysis algorithms. For example, where citation analysis uses the links between particular publications, PageRank calculates rank based on the hyperlinks between pages. For these reasons it seems appropriate to calculate the PageRank

Table 7: Statistics for the different graphs of SIGIR and extended SIGIR corpora

	SIGIR	*Extended SIGIR*	*Document Citation*	
	Co-Authorship	*Author Citation*	*Extended SIGIR*	*SIGIR*
Vertices	1685	4202	2878	766
Greatest Connected Component (GCC)	1415	4055	2701	577
GCC % Connection	83.98	96.5	93.8	75.3
No. Of Clusters	89	130	112	169
Second Biggest Component	15	12	15	4
Diameter (GCC)	18	8	14	12
Average Path Length (GCC)	6.81	3.39	5.39	4.68
Average Clustering Coefficient (GCC)	0.091	0.570	0.627	0.477

of authors and publications within the extended SIGIR corpus.

In constructing the corpus, all citations of SIGIR documents are followed and information is gathered on the citing document. Those documents which are not from within the SIGIR proceeding themselves however, are treated as documents with no citations of their own. While the number of citations made to the documents is recorded, no attempt is made to collect any citation information as shown in Figure 51. This is necessary so as to create a closed and finite data-set.

Following only SIGIR document citations means that the authors of non-SIGIR papers may be seen in the same light as anonymous authors within the Wikipedia context (Adler and de Alfaro, 2007). These anonymous authors are unable to build up a reputation and rating of their own, but do contribute to the reputation of SIGIR authors through their presence. Since these authors have cited SIGIR papers, they have been influenced in some way by the work they cite. Another analogy is that of web forums where un-registered user are considered to be of less value than users registered with the forum. This restriction of following only SIGIR document citations also provides the opportunity to create a complete graph and citation network for authors within the SIGIR proceedings.

4.5.1 Citation Network of SIGIR Publications

The graph of paper-to-paper citations is directed and so the notion of diameter and average shortest path does not apply to it. By removing the directional constraint from the graph however, both measures may be calculated. The removal of direction from the

graph allows co-citing publications to be connected by moving from citing document to cited and then back to a different citing document. Two papers are said to be **co-cited** if they are cited by the same paper (In the context of the threading discussed earlier, these two papers would have the same parent).

A "small world" (Watts and Strogatz, 1998) is likely to exist within the extended SIGIR community, implying that there are a number of papers which are used regularly as references by SIGIR publications. This does notionally make sense since papers from within the SIGIR proceedings would be making use of well-respected and highly-referenced papers. Figures for the SIGIR corpus alone (Table 7) within the same 10-year window show an decreased diameter and average path length of the greatest connected component (GCC), but lesser percentage coverage. The strong intra-conference citing of SIGIR papers means an average path-length of 5 papers can be used to connect any two SIGIR publications. The citation of highly-cited papers from the IR community as a whole however, are more prevalent than those made between SIGIR papers themselves.

The PageRank of papers within the paper-to-paper citation graph has been calculated and is presented in Table 8. While some of the entries within the top 10 are perhaps surprising, having very few citing documents, the citing SIGIR papers are themselves highly cited or these papers are co-cited with other highly ranked papers. This fact compounds the evidence that there is a tight cluster of papers which have been cited highly or have been cited by other SIGIR papers which are highly cited. Coincidentally, many of the authors of these papers are also highly cited within the collection. This could be a consequence of having a single paper within the top 10 scoring papers, or due to high publication rates. This is discussed in the following section.

4.5.2 Citation Network of SIGIR Authors

Social network theory allows us to study the network around authors, viewing the citation of articles as a social interaction. We are therefore interested not just in what authors are saying to each other (i.e. the collaborations which they form), but also what authors are saying about each other. This interaction forms the basis of an implicit social network for each other, where the network of an author is created through the influence of their publications as well as through their collaborations. Influence is measured through analysis of citation patterns for an author. An author may be thought of as influential if their papers are highly cited. (Note there is no differentiation between positive and negative citation. An author is considered to be of influence if their work is cited for any reason. This seems logical as referencing is a conscious decision made with the intended purpose of relating the work of the author to that of the cited work in some way (Garfield, 1997; Salton, 1971a).)

Table 8: PageRank for citation of SIGIR papers within the extended SIGIR corpus

Rank	*Title*	*PageRank*	*Year*	*Authors*	*Citations*	
					SIGIR	*Total*
1	A Language Modeling Approach To Information Retrieval	0.022505754	1998	Jay M. Ponte, W. Bruce Croft	67	175
2	Improved Algorithms For Topic Distillation In A Hyperlinked Environment	0.015772241	1998	Krishna Bharat, Monika R. Henzinger	20	144
3	Learning Routing Queries In A Query Zone	0.01048013	1997	Amit Singhal, Mandar Mitra, Chris Buckley	6	18
4	A Re-Examination Of Text Categorization Methods	0.01016484	1999	Yiming Yang, Xin Liu	32	177
5	Feature Selection, Perceptron Learning, And A Usability Case Study For Text Categorization	0.009516502	1997	Hwee Tou Ng, Wei Boon Goh, Kok Leong Low	9	51
6	The Use Of MMR, Diversity-Based Re-Ranking For Re-Ordering Documents And Producing Summaries	0.009321229	1998	Jaime G. Carbonell, Jade Goldstein	15	75
7	Predicting The Performance Of Linearly Combined IR Systems	0.009022063	1998	Christopher C. Vogt, Garrison W. Cottrell	2	15
8	Exploiting Clustering And Phrases For Context-Based Information Retrieval	0.008708239	1997	Peter G. Anick, Shivakumar Vaithyanathan	3	23
9	Distributional Clustering Of Words For Text Classification	0.008270518	1998	L. Douglas Baker, Andrew McCallum	11	69
10	Variations In Relevance Judgments And The Measurement Of Retrieval Effectiveness	0.007835358	1998	Ellen M. Voorhees	16	38

From the underlying undirected graph of author-to-author citations we can discover the characteristics of the author citation network within the extended SIGIR corpus. Conversely to what was shown by Kirsch (2006) for the co-authorship network of SIGIR for its first 25 years (1971, 1978-2002), the author citation graph of the 10 years of SIGIR conferences between 1997-2007 exhibits what Watts and Strogatz (1998) refer to as a"small world". This echoes the ideas put forward by Milgram (1967) (see Chapter 2). It is interesting that this is the fact. One possible explanation is that the authors who are already established have contributed more to the last 10 years than new-comers. By expanding their collaboration network, and introducing new authors at the same time, established authors have helped to connect up the co-authorship graph of SIGIR. This again echoes the preferential attachment that is in evidence for publications in general, as well as for blogs and posts.

Kirsch (2006) examined the co-authorship PageRank of papers from the initial 25 years of SIGIR proceedings. We calculated the PageRank of authors in the co-authorship network for our extended SIGIR corpus (1997-2007) and find that the top 5 authors (Table 9) more closely resemble those of Hiemstra et al. (2007) as shown in Table 10(a). Looking at the citation and publication information for these authors it is striking that several of the authors began publishing around the mid '90s. When we take this fact into account, it is less surprising that such I.R. luminaries as C.J. van Rijsbergen do not feature. These authors have not published new works as frequently in the last 10 years and so are not cited as often.

Table 11 shows a dramatically different picture. While some of the authors from the co-authorship top 10 also feature in this list (most noticeably W. Bruce Croft retains his number 1 rank), the list is dominated by the authors of language modeling papers. This is hardly surprising as the paper with the top PageRank from the citation graph is a seminal language-modeling paper. Moreover, this paper (*"A Language Modeling Approach To Information Retrieval"*[9]) is far more highly-cited (and by highly-cited papers themselves) than any other paper. The author of this paper is, of course, W. Bruce Croft and Jay M. Ponte. W. Bruce Croft is cited regularly by any language-based IR paper, a fact witnessed by his vastly superiour citation count within our extended SIGIR corpus. Ponte lies just outside the top 10. The interesting inclusions are those from 2 to 4 in Table 11. These 3 authors, Peter Schäuble, Martin Wechsler and Páraic Sheridan are nowhere near as highly cited as the other authors in the top 10. Their inclusion however may be explained by the fact that a paper which they co-authored (*"Cross-Language Speech Retrieval: Establishing a Baseline Performance"*[10]) is cited along with SIGIR papers written by Buckley, Croft and Singhal respectively. This would suggest that while the paper itself is not highly referenced, it is highly connected. This leads to the authors being highly connected as well as being highly co-cited with

[9] http://portal.acm.org/citation.cfm?id=291008
[10] http://portal.acm.org/citation.cfm?id=278459.258544

Table 9: Statistics for the top-10 authors by co-authorship within our extended SIGIR corpus

Rank	*Author*	*PageRank*	*Collaborators*	*Publications*	*Citing Authors*	*Citing Papers*	
						Total (Downloaded)	*SIGIR*
1	Wei-Ying Ma	0.004778386	105	60	219	129(100)	13
2	W. Bruce Croft	0.004165636	46	52	521	670(321)	125
3	James Allan	0.003244352	29	33	238	201(109)	29
4	Zheng Chen	0.002959236	56	35	180	102(82)	11
5	Susan T. Dumais	0.002488424	34	25	260	213(113)	23
6	Jamie Callan	0.002445131	32	40	253	234(145)	37
7	Jian-Yun Nie	0.002181312	37	23	163	116(77)	23
8	David Carmel	0.002013416	21	12	127	95(74)	13
9	Tat-Seng Chua	0.001970828	27	20	47	23(23)	5
10	Charles L. A. Clarke	0.001936583	18	17	151	94(68)	25

Table 10: Top 5 authors in the 30 years of SIGIR proceedings (Hiemstra et al., 2007)

(a) Top 5 Co-Authors

Author	*Co-Authors*
Wei-Ying Ma	54
W. Bruce Croft	41
Zheng Chen	36
James P. Callan	28
Clement T. Yu	26

(b) Top 5 Papers Authored

Author	*Papers*
W. Bruce Croft	44
James P. Callan	21
Wei-Ying Ma	18
James Allan	16
ChengXiang Zhai	16

highly cited authors etc.

Looking at positions 5-10 of Tables 9 and 11, we can see that while the average number of collaborators and publications in Table 11 is fewer, the average citations both by SIGIR and non-SIGIR papers is greater. This would indicate that the authors found within the top 10 cited authors are less prolific in their writing, but the papers which they write are highly cited.

In the context of SIGIR as a surrogate for blogs/posts, we may see these writers as akin to bloggers who write a blog which receives a great deal of comments. One side-effect of this phenomenon in real world blogging however is that some writers have actually turned off the commenting on their blogs as a result of huge numbers of comments. An advantage of incorporating the authority weighting of authors introduced in this thesis in Section 4.6.1.1 is that comments may instead be filtered so that comments by 'authoritative' authors are still allowed, similar to Windley et al. (2007).

4.6 Network measures

Much of the work previously carried out on the corpus of SIGIR papers has concentrated on the co-authorship of papers, comparing the writing of a paper to "knowing someone on a first-name-basis". In the context of a social network this is akin to the network of people whom we meet and interact with regularly. It does not however include those people who speak about us or the people we follow. These people we do not know on a "first-name-basis" but may have regular contact with. In the context of the internet or citation, these interactions are characterised by readers who comment on blogs; other bloggers who reference or create "track-backs" to blogs and citation of another's work. These interactions are not necessarily between people who know each other, but when repeated regularly form the basis for some sort of relationship. The advantages of a social network lie in its ability to measure the connectedness and cohesion of the agents or actors within the network, allowing for clustering of like-minded or "similar" agents.

Table 11: Statistics for the top-10 authors by citation within our extended SIGIR corpus

Rank	*Author*	*PageRank*	*Collaborators*	*Publications*	*Citing Authors*	*Citing Papers*	
						Total (Downloaded)	*SIGIR*
1	W. Bruce Croft	0.016781199	46	52	521	670(321)	125
2	Peter Schäuble	0.014449198	3	2	31	14(14)	5
3	Martin Wechsler	0.014449198	3	2	31	14(14)	5
4	Páraic Sheridan	0.012618852	3	2	15	9(6)	3
5	Amit Singhal	0.011909295	12	8	227	164(99)	31
6	John D. Lafferty	0.010854239	10	10	203	284(138)	73
7	Jaime G. Carbonell	0.010282705	14	7	134	138(57)	21
8	Jade Goldstein	0.010282705	3	3	134	136(57)	21
9	Chris Buckley	0.009618015	14	10	209	180(106)	46
10	Adam L. Berger	0.095784880	6	5	163	117(83)	37

By looking at both co-authorship and citation as forms of interaction, we are able to build a better idea of the community around an author and not just the social network itself. The citation of a publication may be though of as a comment on the work of that author. As such, it is similar to the discussion engendered within a web-forum or blogs where users may create threads of conversation. If we take into account the users who interact with each other, and not just the chosen friendships/relationships which users define amongst themselves, a more complete picture of the network is revealed. Users who interact frequently through a forum, say, may not be aware that it is essentially the same group of users with whom they interact (Fiore et al., 2002). By taking into account all interactions, these dynamics are no longer hidden but may be used to pro-actively suggest new relationships to users.

This same idea is possible within the context of paper citations and may enable the better classification of papers during retrieval or other tasks. By looking at the authors who commonly cite each other's work, clusters of authors are created who may not necessarily have formed a group themselves through co-authorship (Hess, 2006). Nonetheless, this cluster has obvious benefits of working together since each feeds off the others' work. This idea is better illustrated in the context of the evolution or growth of a research topic. By looking at the citing documents and authors, we may see where a research idea has been borrowed from one community or area of research and implemented in another. Not only this, but since we are also observing the authors, no citations need be made between papers for this to be possible. If a sufficiently 'authoritative' researcher were to reference a work, then through their own publications and co-authorship community, we may follow the idea as it morphs or traverses the author network. The analogy within the blogosphere is an author of high social-stature within a particular area who comments or links to a blog posting of another user. This post can help to promote the original author. The difficulty we face with citations is that they happen in blocks; all citations to a paper are temporally ordered, but only to a annual granularity. This fact means that following the temporal ordering of citations in our extended SIGIR corpus is less exact.

In his 1973 paper, Granovetter (1973) showed the power of so-called "weak" interaction to connect a network. By studying the interaction of agents within a community he was able to show that the indirect connections that people make through a friend-of-a-friend can prove powerful. This idea has far-reaching implications in terms of the importance of acquaintances within the real world as well as virtual, engendering both research and commercial opportunities. One popular social networking site, Linked-In[11], enables users to build on the weak ties created between friends and colleagues to extend their professional network of contacts. Granovetter studied the effect of micro-interactions on the macro-dynamics of the network as a whole. This idea may be extended in the

[11] http://www.linkedin.com/

current context to take into account the interaction which take place between authors. By looking at the interaction between two authors in the context of citations, as well as the interactions of each author with other individuals within the network, connections can be made between authors who have not in fact cited each other. Instead, the importance or interest of authors in relation to each other may be gauged through the weak ties formed through actual citations.

A citation creates a connection between citing and cited authors. The tie is different to that of an egocentric network where these ties represent the acquaintances or friendships of an individual as studied by Gilbert and Karahalios (2009). Instead the relationship is between research topics, and implicitly the research of the authors involved. The co-citing of authors, the number of times that authors are cited or indeed publish together, provides a measure of their connectedness. While citations are created for a myriad of reasons, an author who is commonly cited by another author implicitly shares a relationship with that author. The previously mentioned metrics of co-citation and co-authorship are two ways in which to add a weight to the connection between authors.

4.6.1 The Value of Authorship

Another way to weight users relative to each other is by measuring the quality of the information that a user provides to the community. We have developed two techniques which we will use to improve the ranking of documents provided in answer to a user's information need. The re-ranking does not necessarily depend on a query, but may also be used as a means to help guide a user's browsing. In this way we can provide users with information on who the most influential authors/participants in a particular situation are based on their overall contributions to the topic being discussed.

We wish to make use of the networks created between users (in the case of citations these users are the authors) to discover the most influential and informative people. This influence does not necessarily come from the volume of information that a user provides to the network, but may come from the fact that a user promotes or causes conversation. In the context of citations this means that an author has written a highly cited paper, citations being thought of as a form of conversation. In the context of blogging and web forums, this conversation is evident in the messages and comments left by users.

As our basis for quality we take the theoretical basis provided by Zhu and Gauch (2000), with the exception of relevance and availability (see Chapter 2). The main premise of the following two equations is that importance flows from commentator to annotation to document. Citations have already been shown to exhibit all the characteristics of annotations, and so in the context of our extended SIGIR corpus we think

of importance flowing from author to citation to article. If an article is cited by authors who are important or influential within the network of authors, then the value of the article should be increased. We may also say that the value of the actual citation-context (comments in the case of web-forums or blogs) is dependent on its author.

In the following explanations of the 2 techniques developed in this thesis - AuthorRank (A_R) and MessageRank (M_R) - we use the term "Annotation" to refer to the text created by a user when commenting on a document. In the context of the extended SIGIR corpus, this annotation is the citation context which we have retrieved using the methods outlined in section 4.3.1. In the context of the web as a whole, these annotations might be comments on a blog, messages within a forum, or (with future advances in annotation technologies) annotations on any publicly available resource.

Each author receives a score based on the annotations which they have created. AuthorRank then allows us to decide which authors should be considered most expert or most likely to have promoted the supplementary creation of information useful to the user community as a whole. A similar idea is employed by Hotho et al. (2006) to aid in ranking pages tagged by a popular social-bookmarking site. By focussing on the influential authors, and adjusting the ranking around them, users are provided with the most interesting and informative results to a query and/or a better browsing experience. If we then go one step further and focus on the conversations between the top ranked authors, we can find documents which are both most likely to satisfy the user's needs, and which also are most likely to serve as the anchor for informative and insightful annotations.

4.6.1.1 AuthorRank

AuthorRank, A_R, takes into account three different characteristics of an author's interactions with the network; the amount the author writes; the level of interaction that the author has with the rest of the community; and the level of influence which the author has over the conversation being had. These factors are combined within Equation (24). Avg_{wc} , is the average amount (a word count) that the author has written per annotation.

$$A_R = \log(Avg_{wc}) * \{\frac{S_T + \alpha * S_B}{S_{TOT}} + \beta * [\frac{R_T + \gamma * R_B}{R_{TOT}}]\} + \log(Avg_r) * [\sum_{x=1}^{n} \frac{r_x}{e^x}] \quad (24)$$

The central part of AuthorRank, Equation (25), takes into account the cohesiveness of the author by looking at the percentage of annotations which are the start/head of a thread, S, verses those which are replies to other annotations, R. In the context of

our extended SIGIR corpus, this is equivalent to citing a SIGIR article which does not cite any SIGIR articles itself thereby starting a thread of citation, S, or indeed citing a SIGIR article which in turn has cited a SIGIR article etc., R. (Note that it is only SIGIR articles that are taken into account in this ranking, therefore an author who has only published non-SIGIR papers within our corpus will receive a score proportionally to the amount that they have written only.) Annotations are further divided into those annotations that have received replies, and those remaining barren (have no replies). The penalising constants α and γ are applied to annotations that remain barren. This is necessary since an annotation/paper that receives no replies should be valued less than one which is simply the last of a thread. This is also true of an annotation that is the originating annotation of thread compared to one found within the thread. The main reasoning for this is that, like hubs (Kleinberg, 1998), the more interesting an author is, the more conversation they promote. β is the penalising constant in this case.

$$\frac{S_T + \alpha * S_B}{S_{TOT}} + \beta * [\frac{R_T + \gamma * R_B}{R_{TOT}}] \tag{25}$$

We are not solely interested in the accuracy or believability of the information contained within the annotation, more in the catalytic potential to create conversation. This is reflected in the last part of AuthorRank, Equation (26), that takes into account the conversation occurring due to an author's comments. We would like to discern how argumentative or provocative an author is. The average number of responses an author's comments provoke, Avg_r, provides a measure of this. These responses include not just the direct replies to the author's comments, but all replies occurring below a comment within a given thread. Conversation may change and the influence of the author's conversation will diminish the further down the thread we go from this author's comments. To reflect this we distinguish between annotations at different levels within the thread. Equation (26) shows the weighting of all responses r_x at distance x from an author's comment.

$$\log(Avg_r) * \sum_{x=1}^{n} \frac{r_x}{e^x} \tag{26}$$

Again, this measure is independent of the validity or believability of the annotations made by the author but instead reflects the conversational/public appeal of the annotations. It has been noted by Krishnamurthy (2002) that *"the number of comments per post is perhaps the truest and most diagnostic metric of the nature of the communication on a weblog. The posts that are the most insightful or controversial get the most comments."* This has also been shown to be true of in-context annotations (Lanagan and Smeaton, 2007).

4.6.1.2 MessageRank

$$M_R = A_R * \{\frac{2\log M_w}{\log T_w * \log T_a} * [\log T_l - \log M_d]\} + \tau * [\sum_{x=1}^{n} \frac{A_{R_x}}{e^{d_x}}] + (1-\tau) * [\sum_{y=1}^{m} \frac{A_{R_y}}{e^{d_y}}] \quad (27)$$

While the AuthorRank of Equation (24) reflects the global characteristics of each author, Equation (27) gives the MessageRank,M_R, of each particular annotation. This rank is affected by the AuthorRank, A_R, of the author who created it, the replies it receives, the depth at which it is found within a thread, and the AuthorRank of authors involved in the annotation's containing thread.

$$\frac{2\log M_w}{\log T_w * \log T_a} * [\log T_l - \log M_d] \quad (28)$$

The size of annotations in terms of message words, M_w, gives the first indication of its impact. Longer messages are considered more important as there is a greater probability of these messages will stimulate further conversation. We also take into account the number of words, T_w, within the entire containing thread of the annotation. In order to judge the influence of the annotation on its containing thread, the average words length of annotations within the thread, T_a, must be calculated. In the context of our extended SIGIR corpus, the length of annotations is replaced by the length of citation context.

By taking into account the length of the thread, T_l, as well as the depth at which the annotation is found, M_d, increased importance is given to annotations which are found higher (or earlier) in longer threads. Annotations from a thread which contains many entries are considered to be more interesting or important by virtue of the fact that more people are interested in the conversation being had (Fiore et al., 2002; Xi et al., 2004). It may also however be a reflection on the material which is being annotated. This equally validates the assumption that longer threads have held the readers' attention for longer, and are therefore more interesting.

$$\tau * [\sum_{x=1}^{n} \frac{A_{R_x}}{e^{d_x}}] \qquad (1-\tau) * [\sum_{y=1}^{m} \frac{A_{R_y}}{e^{d_y}}] \quad (29)$$

In some contexts, news or discussion forums say, a long thread between just two authors may be thought of as a type of "flame war" where the value of the information provided by the authors involved is likely to degrade as the dialogue continues. We therefore take into account the number of authors found within the thread, as well as who exactly these authors are. By doing so, some notion of the general interest of the annotations may be achieved. To account for topic drift or change of focus, the influence

or strength of the interactions between the author of an annotation and the authors, A_{R_x} and A_{R_y}, of any other annotation in the current thread is proportional to the distance between the two authors, d_x, within the thread (Equation (29)). In this way an author who has replied directly to (or is the direct parent of the annotation in question) is of more effect to the MessageRank than other authors within the thread. We make a distinction between the authors, A_{R_x}, who appear above the author of this message in a thread, and those A_{R_y} who appear below the author of this message within the thread. The authors involved in conversation which comes before this author's comment higher up the thread provide information on the value of this message by virtue of the quality of there own conversation. That is, a conversation which is being had between highly valued authors should be of more interest than that of lesser-valued authors. The value which these authors add to a specific message below them however, should not be the same as those authors who occur below, since those authors and their comments (most specifically direct replies) only exist as a consequence of the message in question.

4.7 An Hypothesis Re-visited

Using the two ranking techniques presented here, our hypothesis is that it is possible to improve the ranking of documents relevant to a user's current information need. While relevant documents can be discovered as a result of classic information retrieval approaches, annotations and threads may be used (as a means of explicit human judgement) to re-rank and improve the ranking of relevant documents. By taking into account the query-independent MessageRank scores for each annotation, we are able to judge the quality of information and citation engendered by any article. By subsequently incorporating this into the overall ranking of papers, papers which are not only relevant to an information need from a text retrieval approach, but also those that have created discussion relevant to an information need may be provided to a user. The approach is similar to PageRank in that it uses the linkage structure created by the annotations to provide a query-independent measure of each author. The novelty however is that when calculating the score for each annotation, the author of this annotation as well as the authors involved in the thread are taken into account. It is not just the links which are considered, but also the creators of these links.

An added benefit of our approach is that the query issued when performing a search need not be as focussed as in traditional information retrieval scenarios. Once a topic has been defined, in terms either of a query or indeed through the browsing history of a user, MessageRank and AuthorRank aim to provide information on important members of the community in relation to the current context. In doing so, users are provided with a guide which is not based solely on a text retrieval algorithm, but which also incorporates the past interactions of authors and users with regards to the topic. This awareness of social interaction and history is one of the foundations of Web 2.0 and

as discussed in Chapter 3, allowing users to benefit from the expertise of others within recommendation systems.

Finally to reiterate the main point of this thesis we believe, as shown by Granovetter (1973), that taking into account the micro-interaction of authors helps improve our understanding of the macro-dynamics of a network of authors as a whole. Specifically we believe that through the use of AuthorRank and MessageRank, we can improve the ranking of documents relevant to a user's information needs. In order to test the effectiveness of our algorithms however, we must first develop a ground-truth against which to compare the performance. In the next chapter we shall detail the collection and creation of this ground-truth.

CHAPTER V

EXPERIMENTAL SETUP II

In this chapter we shall discuss the collection and creation of ground-truth data. Against this ground truth we compare the techniques we have developed in this thesis, as well as other state-of-the-art metrics based on both implicit user-feedback and citation analysis. We established a ground-truth against which all the measurements may be compared, by collecting data from a number of experts located at 3 different universities. Statistical analysis is performed on these rankings to ensure a level of consistency and agreement. Once this has been completed, we compare the rankings provided by our experts to that of the well-known Google Scholar search engine, and other methods widely used in current research practices.

5.1 Creation of a Ground Truth

Our aim here is to test the effectiveness of the algorithms developed in this thesis for improving the rankings of relevant results to a query through the inclusion of author information, as well as the author's social network. We also test our approaches performance against that of other widely-used metrics within the information retrieval field. In order to do this however, it was necessary to create a ground-truth against which to compare. To create this ground-truth we asked 12 expert users from 3 different research groups[1] to provide rankings for documents returned as results for a query. These rankings were then combined into an overall ranking. In the next section we describe how the documents for ranking were chosen, along with the ranking statistics created by our expert users.

[1]Experts were taken from within our own research group in Dublin City University, as well as the information retrieval groups of University College Dublin, and Glasgow University

5.1.1 Document Selection

In previous work on the SIGIR proceedings Smeaton et al. (2003) revealed several re-occurring topics within the conference, clustering the proceedings of the first 25 years of SIGIR into distinct topics. Using this as a starting point, we have identified a new larger set of topics which cover the years (1997-2007) of our extended SIGIR corpus.

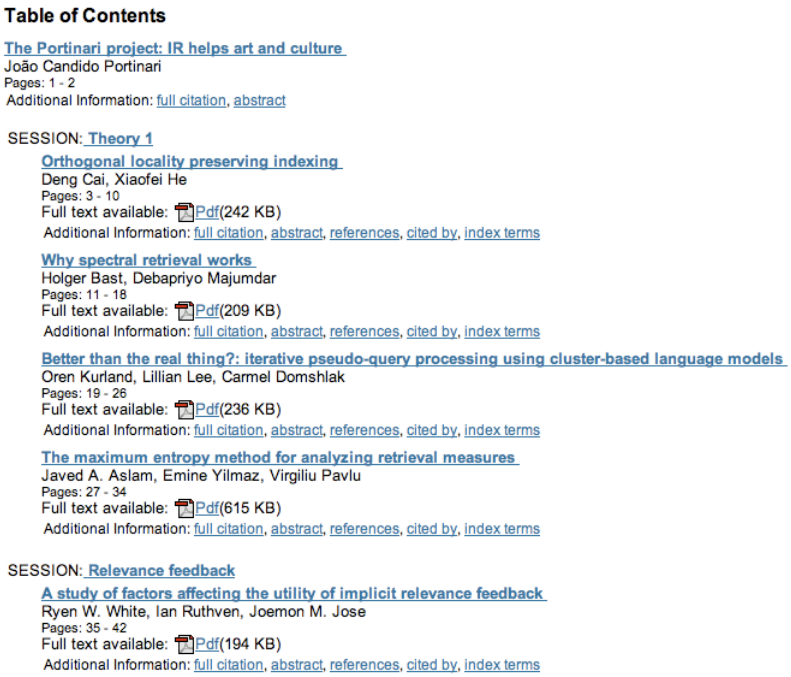


Figure 54: Grouping of papers within the SIGIR proceedings.

The organisers of the SIGIR conference themselves partition the proceedings of each SIGIR conference into different sessions or topics which may be used to help in the manual clustering of documents into conference sessions (Figure 54). By using these session names, and the cluster names from Smeaton et al. (2003) we have chosen 14 topics from which our experts have ranked a selection of documents. We have combined the original topics of Smeaton et al. (2003) with some new ones which reflect the current state of the proceedings; Table 12 shows the topics we have chosen. The number of papers for each topic per year is not constant, since the titles of sessions do not always match exactly. They do however indicate the continuing interest within the IR community for each topic, as well as a new topic interest in the case of, say, spam which has only recently become the subject of more focussed research within the conference proceedings.

For the years 1997-2002 the number of papers for each topic is taken from the clusters created by Smeaton et al. (2003). For the years 2003-2007, the total number of papers is calculated by combining the number of papers within sessions that may be considered part of the overall topic (e.g. *'Web Structure Retrieval'* and *'Linkage Retrieval'* may be considered as sub-topics of *'Linkage Analysis'*).

Table 12: Number of papers per year for each of the topics chosen.

Topic	*Year*										
	97	98	99	00	01	02	03	04	05	06	07
(Collaborative) Filtering[a]	1	1	1	1	2	3		4	3	3	3
Cross-Lingual IR	3	3	1	1	3	4	3	3	3	2	
Distributed IR	1	3	4	2	1	1	3		4	4	3
Document Clustering	2	1		1		3	3	4		3	3
Image Retrieval	1						3	4	3		3
Language Modeling	1	3	1		3	3		4			
Latent Semantic Indexing/Analysis			2	1							
Linkage Analysis		1		3	3	2	3	3	9	7	9
Personalisation[b]		1	2	1				4	4	3	3
Question Answering		1		4	4	1	3		3	3	4
Relevance Feedback	2	1	1	1	1				4	3	
Spam										3	3
Text Summarisation			2	2	3	3			3	3	3
Topic Distillation		1		3	3	2	3				3

[a]Smeaton et al. (2003) cluster title of 'Filtering' has been used for years 1997–2002.
[b]Smeaton et al. (2003) cluster title of 'Users & Search' has been used for years 1997–2002. We have also included User Studies.

The topics chosen give a cross-section of both current research interests and those which have a longer more established history in the context of SIGIR. We can see for example that *'Spam'* has become a more active area of research, while *'Language modeling'* has been less popular in recent years. After analysing the proceedings of the SIGIR conferences between 1997 and 2007, we have selected 14 topics which we feel cast a wide net over the information retrieval topics covered by SIGIR. We then divided these topics into broad and narrow topics, dependant on an advanced search against the Google Scholar website (Figure 55). For a specific query, we noted the number of documents returned for that query within the time period required and restricted to just papers from within ACM SIGIR[2]. Topics were divided into the two groups based on the number of relevant documents returned, a narrow query returning less than 90 documents. which can be seen in Table 13.

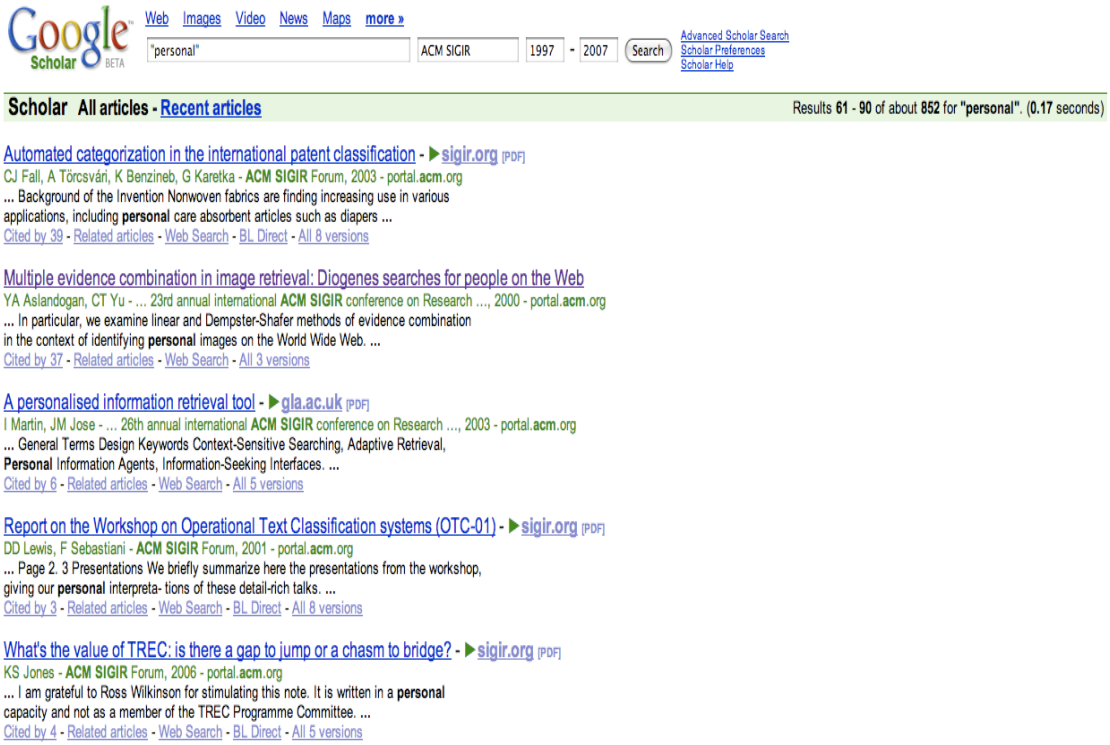


Figure 55: Results for a restricted (or 'advanced search') query performed by Google Scholar.

In order to create a list of documents to present to our experts, we combined the top 30 documents returned from a query against Google Scholar (this query is restricted to the years 1997-2007, and only returns papers from the ACM SIGIR publication list) with a ranked list returned for a query against the citation network of our extended SIGIR corpus. The numbers of documents returned for these restricted queries against

[2]These searches were executed on the 30th November 2008. Google scholar is constantly adjusting its algorithms and weighting features; we have noticed that during subsequent searches of GS we have obtained different rankings for some documents.

Table 13: Number of documents returned by Google Scholar queries restricted to the years 1997-2007, and only the ACM SIGIR publication series.

Topic	*Query*	*Google Scholar Documents*	*Expert Ranked Documents*
Collaborative Filtering (CF)	"collaborative filtering"	68	10
Cross-Lingual IR (CL)	"cross-lingual"	78	10
Distributed IR (DR)	"distributed retrieval"	34	8
Document Clustering (DC)	"document clustering"	97	10
Image Retrieval (IR)	"image retrieval"	99	11
Language Modeling (LM)	"language model"	215	12
Latent Semantic Indexing/Analysis (LS)	"latent semantic"	131	12
Linkage Analysis (LA)	"link analysis"	67	10
Personalisation (P)	"personal"	850	10
Question Answering (QA)	"question answer"	31	9
Relevance Feedback (RF)	"relevance feedback"	350	10
Spam (S)	"spam"	52	6
Text Summarisation (TS)	"text summarization"	83	9
Topic Distillation (TD)	"topic distillation"	53	8

the Google Scholar search-engine are shown in Table 13. In order to perform a search against the publications (as opposed to author) citation graph of our SIGIR corpus, we extended the work of Hiemstra et al. (2007); using the SIGIR abstract file created by them, we manually cleaned and inserted any missing information on author names, titles, and abstracts. We also sanitised the author names so as to conform to the list of authors within the extended SIGIR corpus, repeating the process in Section 4.3.1. Combining this with the PageRank calculations for each document (see Section 4.5.1), we are able to return a ranked list of documents based on the citation PageRank of those documents.

The number of papers given to experts for each topic for ranking may be seen in Table 13. The final list of papers given to our experts was created by combining the two ranked lists, taking the top ranked papers which appeared in both lists. Fewer papers were returned for the queries against the publication citation network; this was a Boolean search against just the title and abstract of papers. As a result, the number of papers which appear in both lists is significantly lower than the number in the Google Scholar list. In cases where the overlap between the Google Scholar and citation publication lists is very low, we augment the final combined list by searching down the two ranked lists, alternately adding the top ranked papers from each list which have not already been added. A threshold is set; if the top 5 ranked papers from each list had been included, the list is complete.

5.1.2 Expert Rankings

Once a list of documents had been created for each topic, expert judgements on the usefulness of these documents to a user were acquired. In order to do this, a number of experts were given a random ordering of the first page of each paper, along with a description of the ranking task. Each expert was given the following scenario:

> A new research student has come to you looking for advice on what papers to read on a particular topic. They have presented you with the papers attached.
> After looking at the front pages of the PDFs, decide upon a ranking of these papers. This ranking should take into account what you judge to be the 'usefulness' and 'value' of each paper to the researcher. This ranking should take into consideration the reading order (i.e. a better paper would be read before other papers). All papers presented to you are assumed to be relevant to the topic.

Experts were asked for an explanation of the rankings that had been provided; what factors affected the ranking of one document higher than another? Experts were also asked to give a rating of 1 to 5 of their knowledge of the topic being ranked;

> 1 - I have had no real exposure to this topic but have rated them to the best of my knowledge
> 5 - I am familiar with this topic and recognise the majority of the authors and/or papers provided.

Table 14 shows the rankings provided by experts for the *"collaborative filtering"* topic. Annotators may be seen to be roughly in agreement on the best and worst papers, while varying more widely on the other rankings.

5.1.2.1 Reasons for Ranking

In ranking the papers which were presented for each topic, experts were asked to consider the ranking with respect to a new research student with little knowledge of the topic. This was done so as to approximate the situation with regards to the annotation and comments in the two Web 2.0 systems of Chapter 3. We wish to show that the use of past users' annotations/comments as a gauge of interest and usefulness for future users is beneficial. By providing the reasoning for their ranking, each expert helps us gain an insight into what factors a human assessor finds important when ranking documents. It

Table 14: Rankings assigned for collaborative feedback documents by the experts

	Annotators						
Title	1	2	3	4	5	6	7
A Nonparametric Hierarchical Bayesian Framework for Information Filtering	4	6	7	7	2	5	5
A Collaborative Filtering Algorithm and Evaluation Metric that Accurately Model the User Experience	2	2	6	4	7	6	2
An Algorithmic Framework for Performing Collaborative Filtering	1	1	1	1	1	1	1
An automatic weighting scheme for Collaborative Filtering	9	7	5	2	8	4	3
Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis	6	4	3	6	4	9	8
Collaborative Filtering with Privacy via Factor Analysis	3	10	8	8	9	7	9
Combining Eye Movements and Collaborative Filtering for Proactive Information Retrieval	10	9	9	10	10	10	10
Effective Missing Data Prediction for Collaborative Filtering	5	8	2	3	5	2	7
Scalable Collaborative Filtering Using Cluster-based Smoothing	8	5	10	9	3	3	4
Unifying User-based and Item-based Collaborative Filtering Approaches by Similarity Fusion	7	3	4	5	6	8	6

was hoped that these factors would coincide with the features used with the AuthorRank and MessageRank algorithms (see Section 4.6.1).

The most common reasons for ranking papers highly were:

- *Author:* The reputation of the author in the publication field, as well as the number of authors was considered very important. An author who had published widely, or published a seminal paper increased the importance of the paper.
- *Institution:* The location of the authors in terms of institution was seen as a good gauge of both quality and influence of the paper. The self-regulation of highly regarded institution provides an effective measure of how useful the publication is likely to be.
- *Year of Publication:* While older papers were sometimes thought to be obsolete, or out of touch, in general experts agreed that older papers provide a good grounding to the topic (especially in the case of seminal papers). New papers were considered useful if they provided a thorough background to previous work, as well as giving a reader a more contemporary view of the field.

- *Content (Abstract, Introduction, Background):* Papers which provided a good overview of current work within the field, as well as seminal works, were given a high ranking.
- *Scope:* Papers which were more general in scope, giving more information on a topic were considered more useful than focussed papers. Applications of specific approaches (to a sub-topic within the topic) were ranked lowly due to the required reading of other more general papers, normally those ranked more highly.

5.1.2.2 Inter-Expert Ranking Agreement

While the reasons that experts gave for ranking documents highly overlapped greatly, the rankings themselves were in no way uniform. Of the 14 experts that rated the topic documents, 6 ranked every topic. Overall we collected 1082 document judgements with an average of 7 judgements per paper. These judgements were given by users with different self-assigned levels of expertise, resulting in a disparity of rank assignments. Table 15 shows the average expertise of the experts who ranked each of the topics.

Table 15: Average expertise of experts who ranked the documents for each topic.

(a) Higher Average Expertise

Topic	*Expertise*	
	Mean	*Median*
Image Retrieval	4.00	3.56
Personalisation	3.00	3.22
Collaborative Filtering	3.00	3.00
Language Modeling	2.50	2.88
Relevance Feedback	2.50	2.88
Link Analysis	2.00	2.67
Document Clustering	2.50	2.50

(b) Lower Average Expertise

Topic	*Expertise*	
	Mean	*Median*
Cross-Lingual	2.00	2.43
Question Answering	2.00	2.43
Text Summarisation	2.00	2.38
Distributed Retrieval	2.00	2.29
Spam	2.00	2.13
Topic Distillation	2.00	2.00
L.S. Indexing/Analysis	1.50	1.83

In order to assess if there is any significant disagreement in these rankings, we have used the Kendall coefficient of concordance (W) to measure inter-rater agreement (Kendall and Smith, 1939)[3]. This measure is explained in detail in Appendix B. It is not possible to use other commonly used measures of inter-coder reliability (such as Krippendorff's Alpha (Krippendorff, 2004) or Scott's Pi (Scott, 1955)), as these assume a nominal dataset, independence of coder's judgements, and lastly an independence of the judgements themselves. Our data is ordinal in nature, and although the experts created the rankings independent of each other, the ranking a document receives is not independent of the other documents.

[3] As with the Wilcoxon Rank-Sum Test used later in this chapter, we have used the implementation provided by the *R* statistical package (R Development Core Team, 2004).

Table 16: Combined expert ground-truth rankings for the *Collaborative Filtering* topic

Title	*Topic*	*Median*	*Mean*
An Algorithmic Framework for Performing Collaborative Filtering	Collaborative Filtering	1	1
A Collaborative Filtering Algorithm and Evaluation Metric that Accurately Model the User Experience	Collaborative Filtering	4	4.125
An automatic weighting scheme for Collaborative Filtering	Collaborative Filtering	4.5	5.125
Effective Missing Data Prediction for Collaborative Filtering	Collaborative Filtering	5	4.75
Unifying User-based and Item-based Collaborative Filtering Approaches by Similarity Fusion	Collaborative Filtering	5.5	5.5
A Nonparametric Hierarchical Bayesian Framework for Information Filtering	Collaborative Filtering	5.5	5.625
Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis	Collaborative Filtering	6	6
Scalable Collaborative Filtering Using Cluster-based Smoothing	Collaborative Filtering	6	6.125
Collaborative Filtering with Privacy via Factor Analysis	Collaborative Filtering	8	7
Combining Eye Movements and Collaborative Filtering for Proactive Information Retrieval	Collaborative Filtering	10	9.75

An important factor to take into account with our expert rankings is that they do not provide any sense of comparability beyond their order (i.e. there is no sense of the first ranked document being some measurable amount better than the second, the second than the third etc.), and so no assumptions can be made about the probabilistic distribution of the data.

5.1.3 A Combined Expert Ground-Truth Ranking

Once inter-expert agreement had been established, the average of these ranks may be used to create a suitable ground-truth against which to test the performance of our own ranking algorithms. The median rank of each paper within a topic is used to create a new combined ranking based on these medians. In the case where the median of two papers' ranks are equal, the mean rank of each paper is used to decide their ordering. In this way we obtained an inter-expert based ranking of the papers in each topic. Again, as an example, Table 16 shows the ground-truth ranking of the *"collaborative filtering"* topic.

5.2 Additional Ranking Sources

We have chosen a number of additional sources to provide rankings for the extended SIGIR corpus data. Included in these is the most frequently used free source for scientific publications rankings, Google Scholar[4]. The sources were chosen in order to compare the state-of-the-art automatic approaches to our combined expert rankings. The two factors which contributed to the selection of the initial set of papers shown to the experts, Google Scholar and the publications citation graph, are both used to calculate a ranking independent of each other. The third source is the download counts per publication as found on the ACM portal.

- **Google Scholar:** The launch of Google Scholar (GS) in late 2004 meant that scholars were provided with a free and extensive source of scientific publication search and citation information. Despite the fact that this resource is free (Butler, 2004), it has been shown to compare well with the performance of paid indexes such as Web of Science[5](Pauly and Stergiou, 2005; Harzing and van der Wal, 2007). One of the largest criticisms leveled against GS is its complete lack of transparency in how it decides upon the ranking of important documents: *"Google Scholar aims to sort articles the way researchers do, weighing the full text of each article, the author, the publication in which the article appears, and how often the piece has been cited in other scholarly literature. The most relevant results will always appear on the first page."*[6] (Figure 56) The size of the citation index, along with the choice of which journals and publications to include in the index created many grievances when GS first became available (Yang, 2006). GS relies on the availability of documents on-line which it may then index, creating a bias towards publications and research areas which have a high web presence. (This is not a problem in the context of our work since SIGIR publications evidently have a strong web presence.) As well as this, the lack of a standardised method of citation and result presentation (removal of duplicates; correction of conflicting characteristics such as publication date; lack of easily available information of a result's publishing document/publisher) means that GS is not very suitable to large-scale bibliographic and citation analysis studies (Yang, 2006). These problems do not affect our calculations since all articles of interest to this research are from the ACM SIGIR conference which is indexed well by Google Scholar.

- **Download Counts:** We collected the download counts for each of the SIGIR papers within our corpus. This figure gives the number of times a paper has been downloaded from the ACM portal page in the previous 12 months. Download counts may be seen in much the same way as click-through data; they give some

[4]http://scholar.google.com
[5]http://isiknowledge.com/
[6]*http://scholar.google.com/intl/en/scholar/about.html*

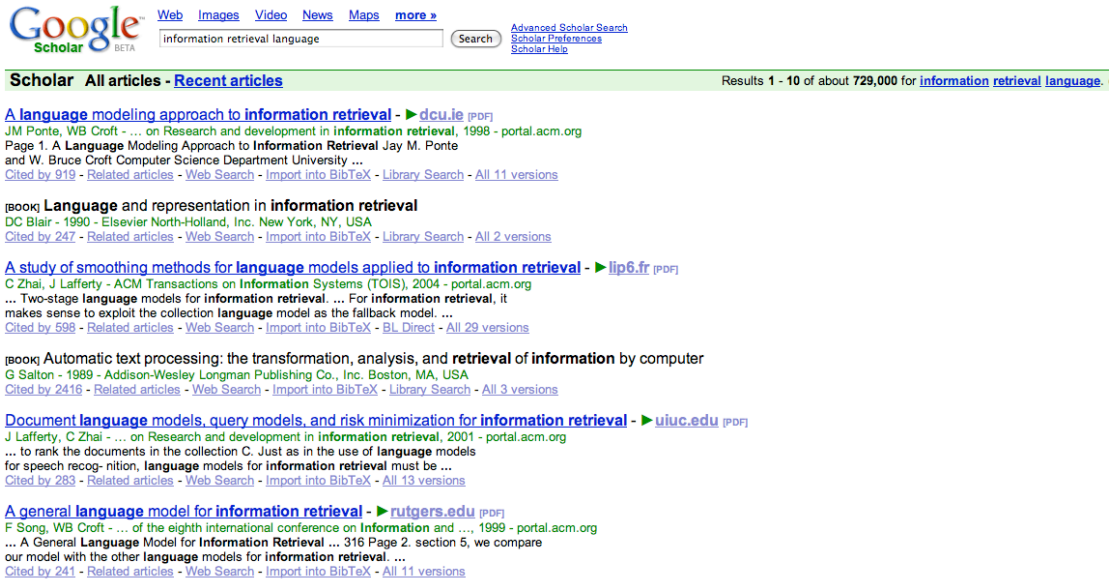


Figure 56: Ranked results as displayed on the Google Scholar results page.

idea of the interest which has been shown in the downloaded article. Joachims et al. (2005) use eye-tracking in conjunction with click data to show the value of implicit feedback in estimating the relevance of search results. They find that, while the feedback is somewhat biased by the presentation format, implicit feedback does correlate well with more explicit feedback in judging relevance. Claypool et al. (2001) look at implicit feedback in the form of reading times and scrolling as a guide to the interest and quality of online resources. In our context, we use the download count as a measure of the interest and value of a publication to the research community. The effort of downloading a paper is in fact greater than that of simply clicking through a search result, and so the download count gives a good estimation of the perceived value of a resource. We can not say that a document which is downloaded by a person will prove useful, but we can say that a person would not bother going to the effort of downloading a publication if they did not see any direct personal benefit in doing so.

- **PageRank(Paper Citation Graph):** The third measure used to rank the papers within each topic was to use the paper citation graph created for our extended SIGIR dataset (see Section 4.5.1). We have calculated the PageRank of each paper and then return the relevant papers ranked by this PageRank. The process used to find contributing papers to the intial set of papers in Section 5.1.1 given to experts is repeated here.

Table 17: Rankings from additional sources for the *"Collaborative Filtering"* topic

Title	*Scholar*	*PageRank (Citation Graph)*	*Downloads (ACM Portal)*
An Algorithmic Framework for Performing Collaborative Filtering	6	1	1
A Collaborative Filtering Algorithm and Evaluation Metric that Accurately Model the User Experience	9	5	3
An automatic weighting scheme for Collaborative Filtering	4	7	5
Effective Missing Data Prediction for Collaborative Filtering	7	10	10
Unifying User-based and Item-based Collaborative Filtering Approaches by Similarity Fusion	3	8	2
A Nonparametric Hierarchical Bayesian Framework for Information Filtering	8	3	9
Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis	2	4	7
Scalable Collaborative Filtering Using Cluster-based Smoothing	1	6	4
Collaborative Filtering with Privacy via Factor Analysis	10	2	6
Combining Eye Movements and Collaborative Filtering for Proactive Information Retrieval	5	9	8

5.3 Comparisons of Rankings

Having performed a Friedman analysis on the rankings provided by each of the measures outlined above, we conclude that there is a significant difference in the manner and outcome of each ranking method. This can be seen in the rankings created for the *collaborative filtering* topic in Table 17. One point to note is that, due to the dynamic nature of the Google Scholar ranking algorithm, we were unable to obtain a ranking of the *personalisation* topic. The restriction of the result list obtained from GS, combined with the changing implementation of the algorithm meant that fewer documents from within the SIGIR corpus were returned[7].

The PageRank citation graph created from our extended SIGIR corpus suffers from one major flaw or weakness; the rankings returned are strongly influenced by the age of the document. This effect is not surprising due to the fixed time-frame and size of the corpus. Ranked lists returned by this method were seen to be roughly chronological in

[7]It was noted that during successive re-issuing of the topic queries against GS, the ranked position of some papers was seen to change. This resulted in some papers disappearing from the ranked list, and others moving up/down the list. In some cases, most notably in the the case of the top ranked document in the *collaborative filtering* topic, the change of rank brought the ranking more in line with that obtained from our experts.

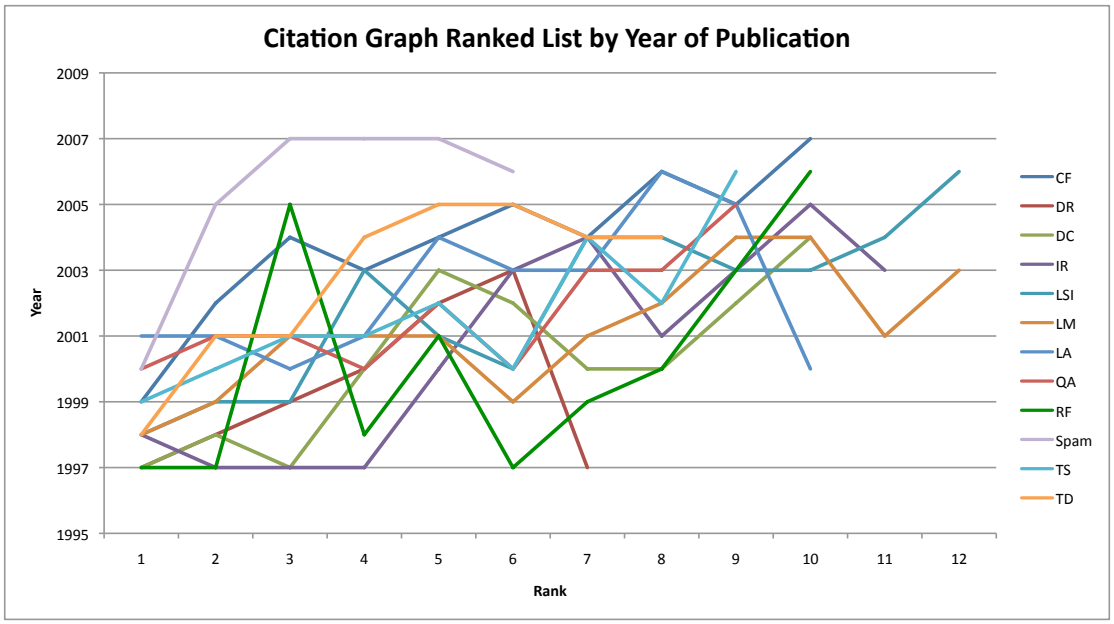


Figure 57: The rankings per-topic returned by the PageRank citation graph may be seen to be roughly chronological in nature.

nature, as shown in Figure 57. There was also positive correlation (0.42 - 0.95) between the per topic rankings provided by the extended SIGIR citation graph and papers ranked by citation counts within Google Scholar. While the citations within GS may come from any paper, citations within the extended SIGIR graph may only come from other SIGIR papers to have any influence on the ranking. This is because (as explained in Chapter 4) only the citations of SIGIR papers are used as edges within the extended SIGIR corpus graph. Even so, correlation is strong and we may conclude that the interest a paper receives from within the SIGIR proceedings is a good indicator for the level of interest within the scientific community at large.

Download counts are used as a measure of explicit interest within the scientific community, downloading a paper being an indication of a reader's interest in the paper. Correlation between the rankings provided by GS and our experts, and a ranking based on download counts of each paper per topic revealed mostly weak to no correlation between the rankings. This may be seen in Figure 58. There is however significant differences in the level of correlation between the download-Scholar rankings and download-expert rankings. This was established through the use of a Wilcoxon Mann-Whitney (Wilcoxon Ranked-Sum) test (Lehmann and D'abrera, 1975) which showed a U-Statistic of 51, and p-value of 0.3804. This leads us to believe that while there is very little correlation between the rankings, the effect of inclusion within the top-ranked papers for a Google Scholar query is not negligible. There is significantly better correlation between the download and scholar rankings which leads us to believe that people are more likely to download a paper which they feel is recommended by an 'authoritative' source. This is interesting in the light of the observations which follow.

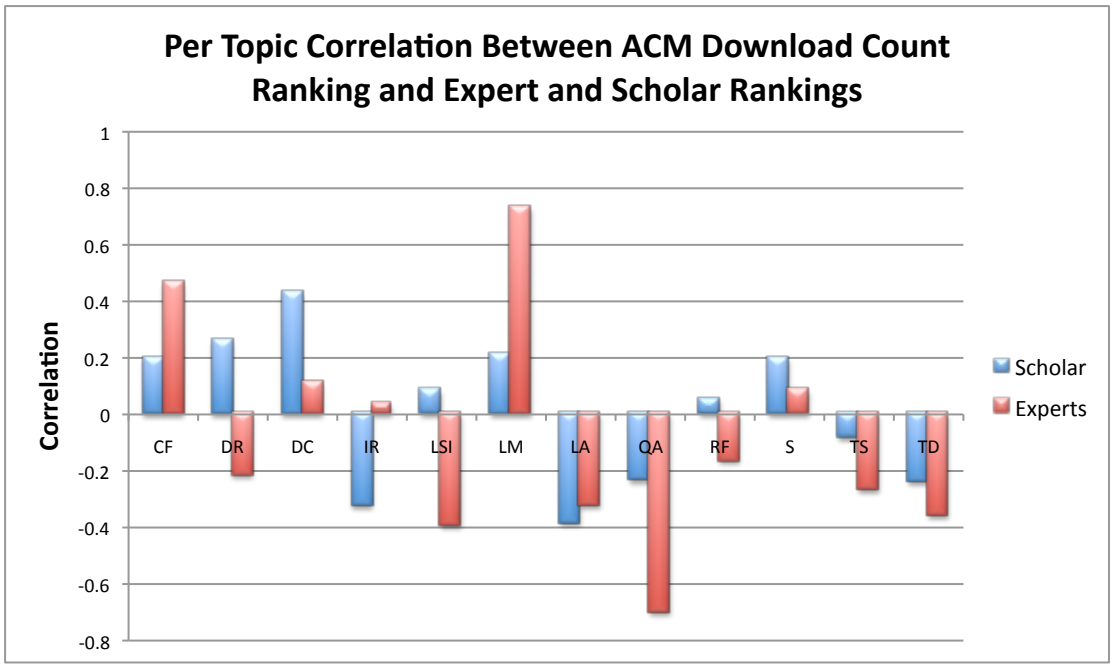


Figure 58: The per-topic correlation of ACM download counts vs. expert and Google Scholar rankings.

5.3.1 Comparing Google Scholar to Our Experts

Google Scholar is the most widely used free source of ranked publications within the scientific community. As stated within its own description: *"Google Scholar aims to sort articles the way researchers do, weighing the full text of each article, the author, the publication in which the article appears, and how often the piece has been cited in other scholarly literature. The most relevant results will always appear on the first page."* Correlations between the rankings provided by our experts, and those created by Google Scholar lead us to believe that the researchers which GS is modelling are those with a basic understanding of the field being ranked. That is, much the same way as a researcher who has little explicit knowledge of the research field, GS seems to use the statistics which are available to it through direct analysis of the papers within the list, and perhaps the authors who have written these papers. It is not possible for the ranking algorithm to take into account past experience or other papers by authors that do not appear within the list. These other papers are therefore considered irrelevant to the task at hand. Quite the opposite to this,when ranking papers human experts will take into account past experience and prior knowledge - prior knowledge which increases with the level of self-assigned expertise of the ranking expert.

This phenomenon may be seen in the decreasing correlation between per-topic expert rankings and GS rankings as the average (mean) expertise of the experts increases, as shown in Figure 59. The purpose of issuing broad and non-specific queries to GS (as shown in Table 13) is to mimic the notion of the novice and inexperienced user who

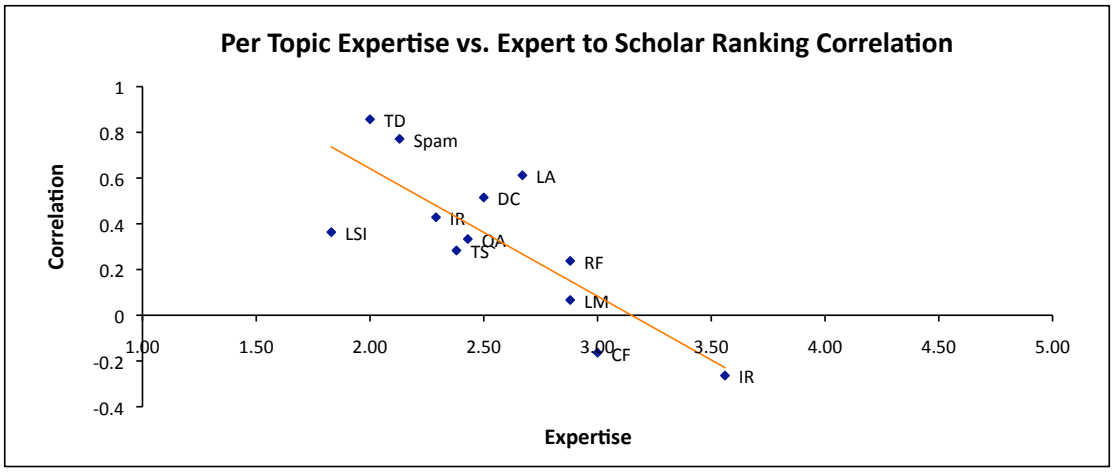


Figure 59: The correlation of per-topic expert and scholar rankings, shown to decrease as average expertise increases.

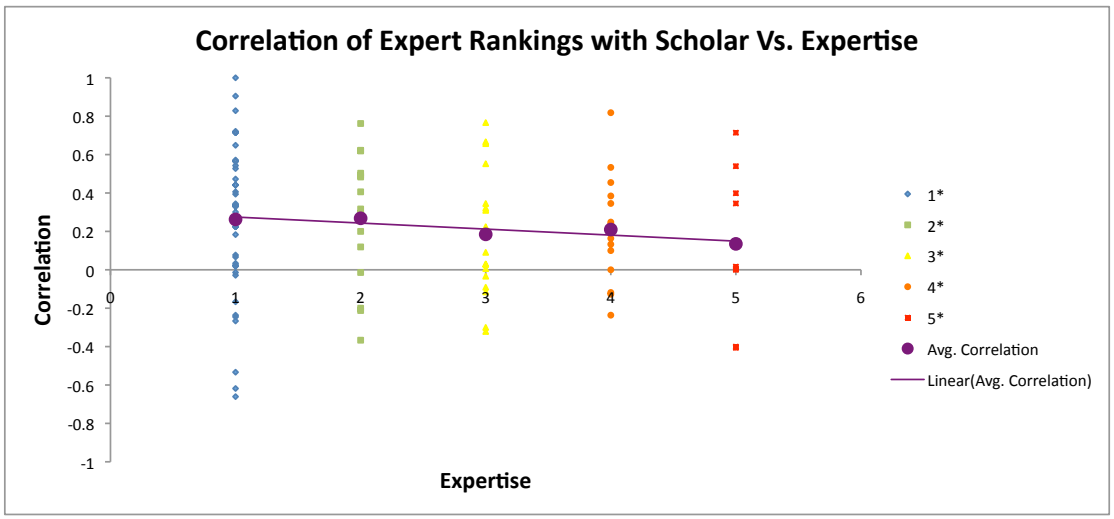


Figure 60: The correlation of per-expert and scholar rankings, divided into differing levels of expertise.

comes to our experts with a selection of papers and no clear idea of the order in which to read these papers. As such, we would have expected the rankings to correlate well with each other, regardless of expertise. We have discovered however, that there is in fact a -0.7922 correlation between the rankings as expertise increases. This leads us to the following conclusion; the rankings provided by Google Scholar are most similar to those provided by experts who have little expertise in the area and can bring no prior knowledge to bear on the ranking.

If we now look at the per-expert correlations with the Google Scholar ranking as shown in Figure 60, we can see that while the correlation is not as strongly negative as on a per-topic basis it is still negatively correlated. The graph presented does not however utilise the within-topic agreements of rankings amongst the experts, but instead

Table 18: Kendall's W and significance levels for per-topic inter-expert ranking agreement by self-assigned expertise level.

Topic	*Papers (n)*	*Experts (k)*	*Expertise*	*Kendall's W*	$\chi^{2(n-l)}$	*p-value*	
Collaborative	10	2	$1^* - 2^*$	0.709	12.8	0.1740	
Filtering		2	$4^* - 5^*$	0.86	15.5		0.0783
Distributed	8	5	$1^* - 2^*$	0.491	17.2	0.0163	
IR		1	$4^* - 5^*$	-	-		-
Document	10	4	$1^* - 2^*$	0.535	19.3	0.0231	
Clustering		1	$4^* - 5^*$	-	-		-
Image	11	3	$1^* - 2^*$	0.222	6.67	0.7560	
Retrieval		5	$4^* - 5^*$	0.352	17.6		0.0621
Language	12	4	$1^* - 2^*$	0.583	25.7	0.0073	
Modeling		2	$4^* - 5^*$	0.811	17.8		0.0852
Latent Semantic	12	5	$1^* - 2^*$	0.423	23.3	0.0162	
Indexing/Analysis		1	$4^* - 5^*$	-	-		-
Linkage	10	4	$1^* - 2^*$	0.471	17.0	0.0493	
Analysis		2	$4^* - 5^*$	0.57	10.3		0.3300
Question	9	5	$1^* - 2^*$	0.656	26.2	$9.55e^{-4}$	
Answering		2	$4^* - 5^*$	0.675	10.8		0.2130
Relevance	10	4	$1^* - 2^*$	0.565	20.3	0.0159	
Feedback		3	$4^* - 5^*$	0.593	16.0		0.0665
Spam	6	6	$1^* - 2^*$	0.378	11.3	0.0452	
		0	$4^* - 5^*$	-	-		-
Text	9	5	$1^* - 2^*$	0.411	16.4	0.0367	
Summarisation		1	$4^* - 5^*$	-	-		-
Topic	8	5	$1^* - 2^*$	0.579	20.3	0.0051	
Distillation		1	$4^* - 5^*$	-	-		-

looks only at the level of agreement between each self-assigned expertise level's ranking and that of Google Scholar. It is interesting non-the-less that the divergence of expertise and GS is repeated at this level also.

To further study this observation at a topic level, we have broken the experts up by level of expertise, again using the Kendall's W measure for agreement. We have measured the agreement between experts of expertise level $1^* - 2^*$ (being experts who feel they have below average, somewhat lacking expertise of the area), and those of expertise level $4^* - 5^*$. These measures of expert agreement may be seen in Table 18. From this we can see that the agreement between experts of level $1^* - 2^*$ is (with the exception of *collaborative filtering* and *image retrieval*, two topics that have the least number of expert rankings) universally significant at a p-value of 0.05. For expertise level $4^* - 5^*$ significance is achieved at a p-value of 0.1. The W measure is universally greater for the more expert raters. One possible explanation for this is that the factors which influence those with less expertise are less well defined than those which influence those

with greater expertise. As a result, the agreement between raters with lower expertise is on average lower.

There are fewer experts who have rated themselves highly in each topic; in some cases there is just one expert. For this reason we have fewer significant $4^* - 5^*$ expert ranking agreements. Figure 61 shows the correlation levels between the levels of expertise and Google Scholar rankings on a per-topic basis. We see that the correlation between the rankings provided by lower expertise levels and GS is higher. This difference is significant as shown by a Wilcoxon rank-sum test, showing a U-statistic of 34 and a p-value of 0.0539. In the topics where just one $4^* - 5^*$ rated expert is present, we have used the ranking provided by the highest rated-expert to see if this affects the significance of expertise-level. The difference is still significant, showing a U-statistic of 61 and a more pronounced p-value of 0.2110. This however may only be taken as anecdotal evidence of increasing divergence of the GS and expert rankings.

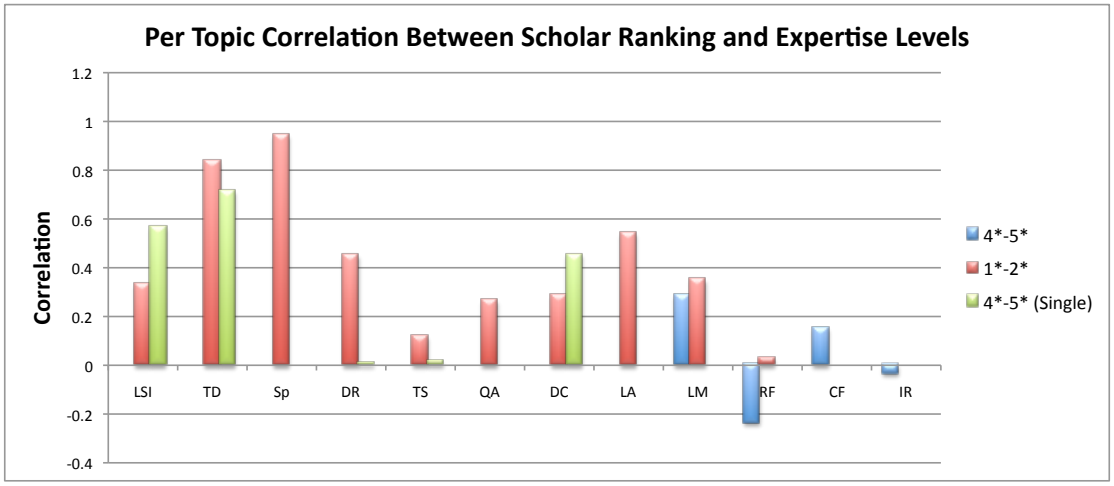


Figure 61: The correlation of per-topic expert and scholar rankings, divided into differing levels of expertise

5.3.2 Harnessing Community Expertise

While it may be argued that the ranking Google Scholar provides is designed to best fit user expectation and therefore need, we do not feel that this ranking is an optimal ranking. By effectively simulating the rankings provided by a more novice rater or expert, GS provides a ranking which is perhaps closest to the expectations of a novice query-issuer. The ranking returned is designed to be closest to one the user issuing the query would create themselves. If the person issuing the query has little knowledge of the area, in order to create a ranking they would have to rely on indicators such as author, conference, year, number of citations, institution etc. to provide a measure of publication importance. The importance of each of these factors is influenced by the prior knowledge that the person has.

In order to create a ranking that more closely reflects an experienced rater or expert would create, we must attempt to harness the expertise and prior knowledge of these experts. To do this we chose to view their interaction with the community as a whole as an indicator of their understanding and experiences within a particular area. These interactions are modelled by the algorithms which are discussed in Chapter 4, *AuthorRank* and *MessageRank*.

Before combining the different features of users' interactions into these algorithms however, we must look at the impact of each of these features alone. In the next chapter, we shall examine the ability of each feature to independently replicate the behaviour of our expert users in creating a ranking of documents for each of our topics. Once we have done this, we shall look at optimal methods for combining the features in order to take full advantage of each of their strength while minimising the weaknesses.

CHAPTER VI

EXPERIMENTS

In this chapter we detail the experiments which have been undertaken to prove the effectiveness and usefulness of the algorithms detailed in previous chapters.

Firstly we describe the systems we have built to perform our experiments, gauging the performance of individual elements of our algorithm in creating a ranking correlated with our experts'. We then combine these features to take advantage of each of their distinct characteristics.

We shall also analyse the effectiveness of current state-of-the-art citation analysis algorithms in the SIGIR and Web 2.0 context. Citation analysis allows us to use current techniques to provide a measure of importance to an author or user. We then look at ways in which to measure the importance of a citation-context as an individual datum.

Finally, we show that the techniques we have developed and trained on our extended SIGIR corpus are indeed of benefit in improving the rankings of documents returned in response to a query.

6.1 A Search System

We would like to look at the impact of each of the features of a user's community interactions that we have identified based on the work of Fiore et al. (2002) and Zhu and Gauch (2000). In order to do so, we must first retrieve a list of relevant documents from our corpus. To do this, we have created two search systems which utilise the Lemur Toolkit (Allan et al., 2003) to build indexes against which to search. These two indexes are made up of the documents from within our SIGIR corpus, but differ in one major respect:

- **SIGIR_Txt:** This index is made up of the documents from within our SIGIR corpus. We have only included the documents which are from within SIGIR proceedings, disregarding the referencing non-SIGIR documents. These documents are considered non-relevant to any queries that we will issue, being from sources external to the SIGIR proceedings. In our previous ranking experiments, experts were only given papers from past SIGIR conferences, and so any paper from a source external to SIGIR is judged as irrelevant.

- **SIGIR_Comb:** This index contains the same restricted subset of our original extended SIGIR corpus as described above; however here we have added all direct citation-contexts to each SIGIR paper's text. Each document now consists of the text from the original SIGIR paper, plus all text from citations made directly to the paper by other papers. A recognised weakness of our corpus is that (unlike the SportsAnno and Annoby corpora) some of the citation-contexts (those from SIGIR papers which have cited SIGIR papers) come from within other documents which are themselves contained within the index. This however is not seen as a great problem as each citation-context, or annotation, is given its own ID and is therefore seen as an annotation in its own right.

With the two indexes *SIGIR_Txt* and *SIGIR_Comb* available, we may now begin retrieval of potentially relevant documents, as described in Chapter 2. We have chosen the TF-IDF implementation within the Lemur Toolkit (Zhai, 2001) as a basis for ranking the documents returned in answer to a query. We have chosen to use the TF-IDF method as this is a standard method within the field of text-retrieval. For a more in-depth explanation of its origins, see Section 2.1.2.2. This implementation uses a document-length normalisation approach as specified by Robertson and Walker (2000):

$$tf_d = \frac{(k_1 \times tf_i)}{tf_i + k_1((1-b) + b \times dl/avdl)}$$

tf_i = Term frequency measure of term i in document d
$k_1 = 1$
$b = 0.5$
dl = Document length of document d
$avdl$ = Average length of documents in the corpus

This ranked list may at times be many hundreds of documents long, with the relevance of a document to the query becoming negligible the lower down the list it is found due to the nature of our corpus. We are interested in the power of the distinct and combined author features to re-rank and improve the position of the specific documents for which we have judgements. These are the 6-12 documents for each query that have

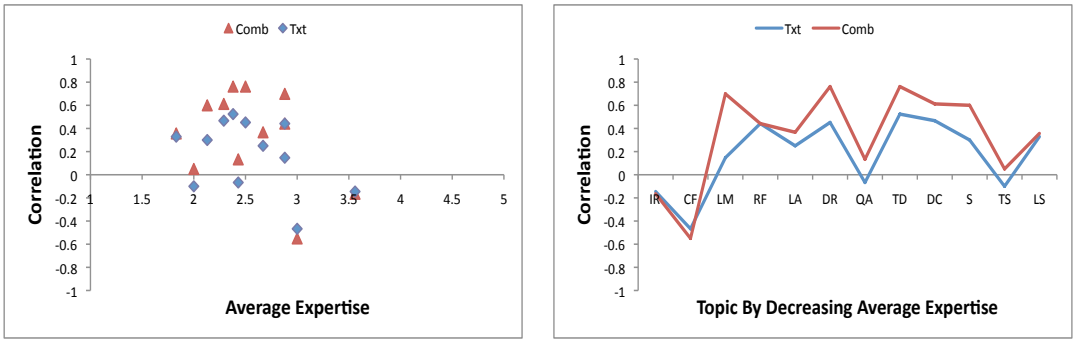
received rankings from our experts. Since these documents are universally returned within the top 100 documents, we have truncated our ranked lists at the 100^{th} document. This also seems reasonable considering that searchers will rarely look beyond the top 5 or 10 documents returned for a query (Silverstein et al., 1999). Also, the inclusion of lower-ranked documents into the re-ranking set is likely to produce more noise than benefit.

6.1.1 Document Relevance

The 12 queries issued to Google Scholar (Table 13, pg.132) to obtain the documents for our expert rankings were issued against each of the two Lemur indexes. For each query, Lemur returned a ranked list of documents it believes to be relevant/match the query. As stated above, we have limited this list to contain just the top 100 ranked documents. We would like to see the rank positions of those documents that have been ranked by our experts (e.g. in the case of a search for the terms 'language modelling', Lemur will return 100 ranked documents, but we are only interested in the 12 documents previously judged and ranked by our experts). We refer to these $\sim$12 documents alone as the 'relevant' documents; in all the results which follow, we have based our calculations of rank correlation and average precision on the documents which we asked our experts to rank. In terms of average precision (AP), this means that it is calculated on the ranks received by the expert-viewed documents. We have calculated average precision so as to gain some idea where the relevant documents have been placed in the rankings created; high correlation with the expert ranking coupled with low average precision, for example, shows that whilst the documents have been ordered in a similar way to that of the experts' ranking, the documents have been found lower down in the ranked list. We use the Spearman Rank Correlation, ρ, for all measures of correlation that follow.

6.1.2 A TF-IDF Baseline

Before looking at the effects of author features, we must calculate the effect of citation-context inclusion on the TF-IDF baseline. This is the baseline ranking from where we take the top 100 documents and perform any re-ranking. Figure 62 shows the correlation of the TF-IDF ranking from each of the indexes with the expert rankings. These are the per-topic correlation figures, showing how well each of the baseline TF-IDF rankings for documents correlated to the rankings provided by our experts. We can see that the inclusion of the citation-contexts significantly improves the correlation ($p = 0.004$). AP is also significantly improved from 0.65 to 0.7 ($p = 0.03$). This result is not surprising and is in agreement with the findings of Ritchie et al. (2008); the inclusion of citation-contexts provides useful index terms and aids in the ranking of documents. Lemur does perform document-length normalisation, so this improvement is not simply due to the



(a) (b)

Figure 62: Comparisons of the baseline TF-IDF rankings from the two Lemur indexes with the experts' rankings

fact that larger, more heavily-cited documents have a higher probability of appearing at the top of the list.

Two things are immediately noticeable from the graph in Figure 62(b); the correlation between the TF-IDF baseline ranking and our experts' ranking for 'language modelling' (LM) is far higher in the combined index SIGIR_Comb; the correlation of the TF-IDF baseline with the topic rankings of higher average expertise is very poor. The first of these points may be explained by looking at the documents which are of relevance for the language modelling topic. The top-ranked document (*'A Language Modeling Approach to Information Retrieval'*) is by far the most cited paper in the corpus. Also, looking through the citation-context text, it is immediately obvious that the term 'language model' is particularly prevalent. Since these are the keywords against which we are performing our search, the inclusion of such citation-contexts will undoubtedly lead to a higher ranking for the relevant papers.

The second observation on the performance of TF-IDF in relation to higher expertise rankings shows us that when ranking papers in order of importance, experts look at features external to the text of the document itself. This observation is in keeping with the reasons given by experts for ranking one paper higher than the next. Since TF-IDF only takes into account the actual text content of a document, and subsequently has no knowledge of author reputation, institution, citation history etc., the ranking it provides is more basic. It is more akin to that created by a novice user who again has little background knowledge of the topic/query that generated the document list to be ranked. We can also see that the correlation of TF-IDF ranking to expert ranking on the *"question answering"* (QA) topic seems anomalously low. One possible explanation for this might be that, although the rankings provided by the experts in general for this topic were significantly correlated (see Table 18, pg.143), the rankings provided by those with higher expertise were not. This same point may explain some of the lower correlations found for *"link analysis"* (LA) later, though both observations are

speculative.

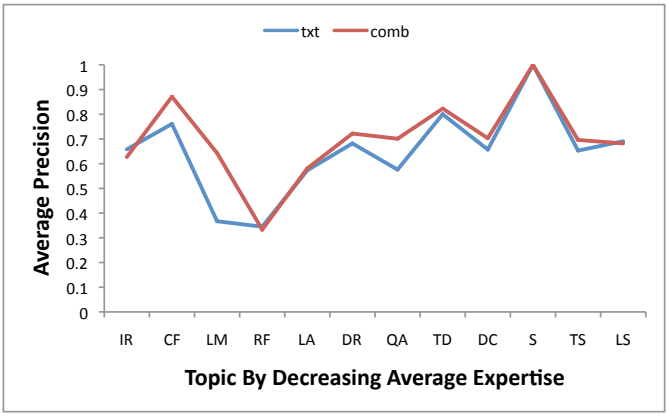


Figure 63: The average precision of TF-IDF for the two Lemur indexes in finding the relevant documents

The average precision achieved by the TF-IDF ranking (Figure 63) on each system is 0.65 and 0.7 respectively. *"Spam"* (S) has an AP score of 1.0. This remarkable result may be due to the number of relevant documents in total within the corpus being lower than for any other topic, as well as the fact that the documents chosen for expert ranking came from both Google Scholar and the eXist databases top ranked documents. This was not generally the case. A consequence of this is that, unlike in TREC, the documents we consider relevant may not always be the top ranked documents [1]. The purpose of reporting the AP figure here is to show the effect of re-ranking procedures on both the ranking of our relevant document in relation to each other (the correlation with expert ranking), as well as the effect on the overall ranking (the average precision). Again we see that language modelling is most affected by the inclusion of citation-contexts.

6.2 Author Value Revisited

We would like to see the effects of both message and author attributes on the correlation and ranking of papers within our extended SIGIR corpus. Before doing this, we must first look at the contribution that can be made by current state-of-the-art measures. In terms of an author's contribution to the importance and relevance of a paper, we first look at the effects from inclusion of the h-index, g-index and m-index of an author. As stated in Section 4.1.2.1, the g-index and h-index help to show the contributions of an author to their field of research. The m-index looks at the overall impact of the work.

In order to discover the contributions made by these measures, we must first combine them with the rankings provided by the Lemur TF-IDF scores. This is done using a

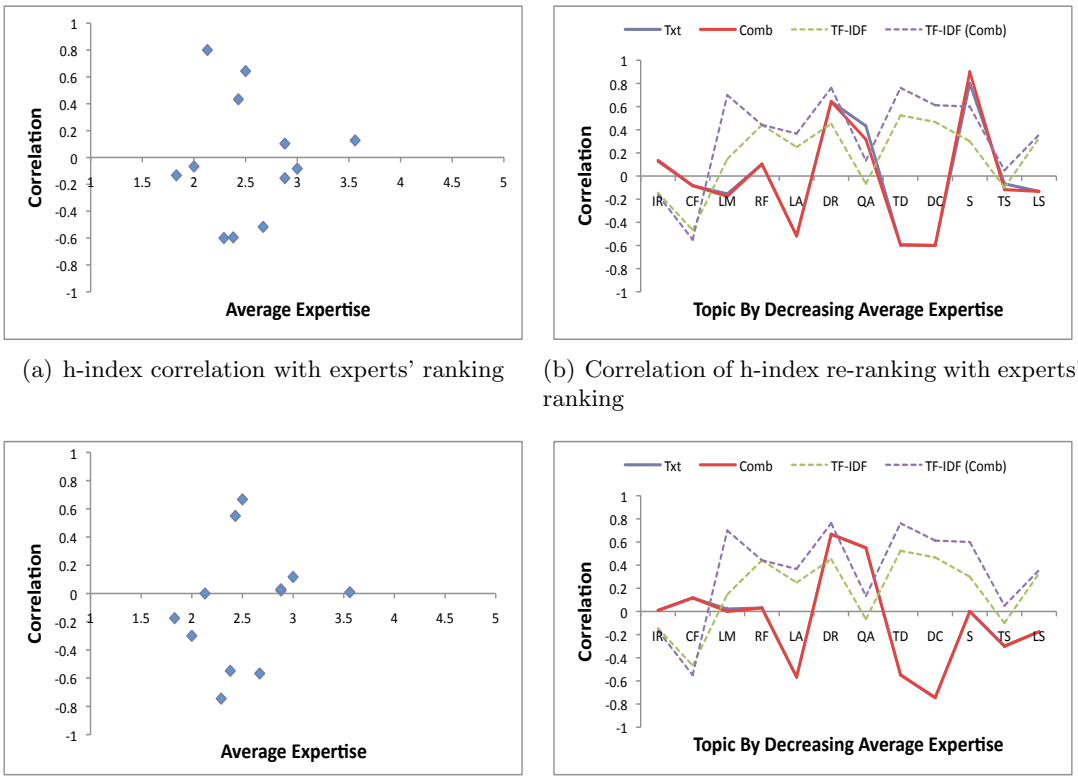
[1] Remember that for a document to be ranked by our experts, it should first appear highly in the rankings of both Google Scholar and the eXist database. Documents after this were chosen alternately from either ranked list.

linear combination of the form:

$$Score_{doc} = \alpha \times Score_l + \beta \times F_x \tag{30}$$

where $Score_l$ is the Lemur TF-IDF score for a document, F_x is the score given by some calculated feature x, and α and β are normalised weights in the range $[0, 1]$. The weights sum to 1, and we use an exhaustive grid-search algorithm to find the optimal weights for each feature.

We choose to use the average of all authors' feature scores as the score given to a paper, since there is no obvious way to weight each authors' scores that gives maximum importance to the most important author.[2]



(a) h-index correlation with experts' ranking

(b) Correlation of h-index re-ranking with experts' ranking

(c) g-index correlation with experts' ranking

(d) Correlation of g-index re-ranking with experts' ranking

Figure 64: Comparisons of the g-index and h-index re-rankings of the TF-IDF baseline from the two Lemur indexes with the experts' rankings

Figures 64(a) and 64(c) show the correlation achieved when the initial list of documents returned by the Lemur TF-IDF scoring function are treated as a random set of

[2]While it may seem obvious to only take the score of the first author into account, this does not give the full picture of a author's importance. In the case of SportsAnno and Annoby a comment is made by a single author; papers in SIGIR are often published by more than one author. In order not to create a bias to papers authored by many people, we do not use the sum of all authors' scores. In research, papers are frequently published by a student, and another author; their supervisor. If we take only the student's score, we may miss the fact that the paper was co-authored by a highly cited researcher.

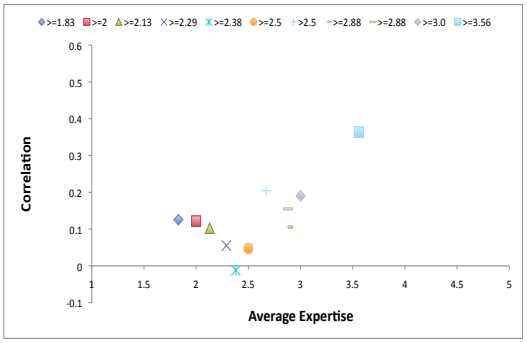
documents; this random set is then ordered by the average g-index and average h-index of the authors of each paper respectively. We take the average of the respective indices for each author of a paper. Attributing this average to the related paper, we then re-rank papers based on this new score. We can see that the effect of both author measures is positive for the correlation with the expert rankings of highest expertise; the effect overall however is shown to be detrimental. The g-index does perform slightly better on average; this may well be due to its attempts to improve on one of the h-index's greatest flaws: As stated, h-index does not in any way give credit to publications by an author which are in some way 'seminal', or very highly cited.

As we have shown in Section 5.3.1, the rankings provided by people of greater expertise as are markedly different from those of lesser expertise. For this reason we do not look at the performance of our different measures simply across topics, but instead choose to split our topics into two classes. These classes are delimited by the level of average expertise of our experts. From Figure 65 we can see that the performance and correlation between the expert rankings and the rankings created by the different features follow a distinct pattern.

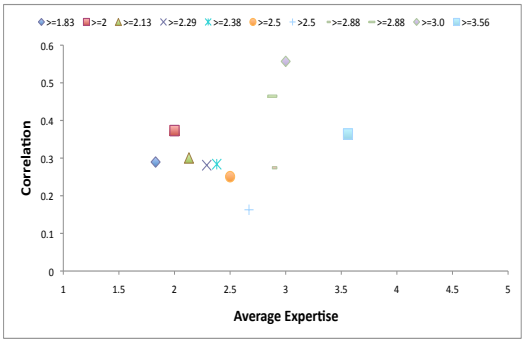
In order to create the data-points within each of the plots of Figure 65, we have looked at the average correlation between the experts' ranking and that of the each of the features' ranked lists. For each point in, say Figure 65(a) (the process is repeated for all features), we compare the correlation of the experts' ranking with the rankings for the topic of highest expertise. We then repeat the process, but this time include the rankings of the topic with the second highest expertise, averaging the correlation of highest and second highest expertise. This process is repeated until all the topics are included. By doing this we are able to see the correlation averages for all the topics. As we can see, there are two quite distinct groupings or classes of correlation; there is a class of topics which have lower expertise (topics with average expertise equal to or less than 2.5); and those topics with a higher average expertise (over 2.5).

We would like the rankings created by our measures to more closely approximate the rankings which a more expert user would create. In other words, we would like to maximise the correlation of our measures' rankings with that of the expert ground-truth for topics which are of higher expertise. From the data we can see that to do this, we should be looking for weighting combinations which maximise the correlation for topics with an average expertise higher than 2.5. We are not interested in the correlation figures for those topics with average expertise equal to or below 2.5, though it would of course be best to maximise these correlations also. This however is not our primary goal.

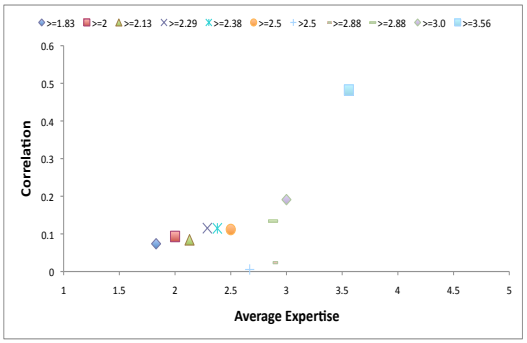
If we look at the topics in terms of two disjoint classes (those topics ranked by experts with an average expertise over 2.5, and those ranked by experts with an average expertise



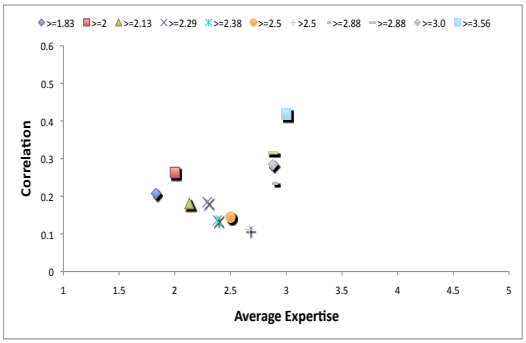
(a) Classes for 'start threaded' on SIGIR_Txt index



(b) Classes for m-index on SIGIR_Txt index



(c) Classes for 'replied threaded' on SIGIR_Txt index



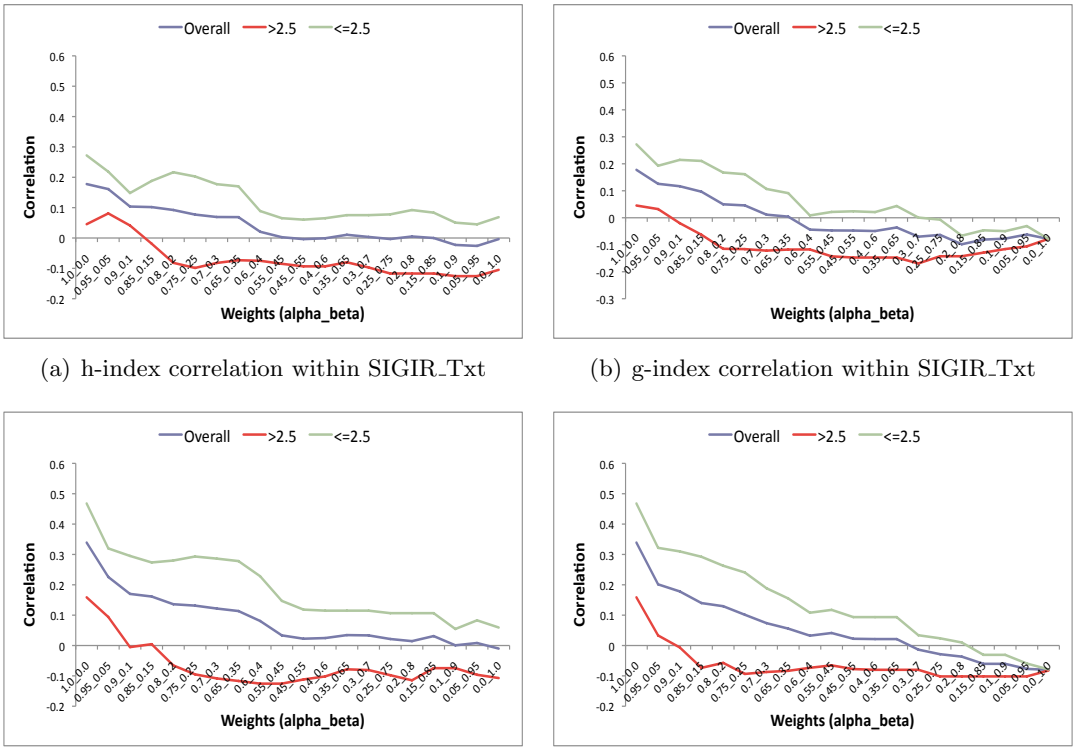
(d) Classes in 'responses' on SIGIR_Comb index

Figure 65: The division of average expertise into two distinct and disjoint classes

equal to or below 2.5), we see a slightly different picture. This is shown in Figure 66. In the case of the SIGIR_comb index (containing citation-contexts) the inclusion of the author measures still has a detrimental effect on the over correlation with the expert rankings. The highest correlation is achieved with $\alpha = 1.0$ and $\beta = 0.0$, ignoring g-index and h-index measures completely. With the SIGIR_Txt index a slight improvement in average correlation within the higher expertise topics is achieved by setting $\alpha = 0.95$ and $\beta = 0.05$, but this is not significant ($p = 0.234$)[3]. AP scores, while approximately linearly decreasing, are also improved within the more expertise rankings; inclusion of h-index information significantly increases AP ($p = 0.08$) from an average precision of 0.54 to 0.65; inclusion of g-index information yields a significant increase ($p = 0.07$) of 0.56 to 0.65. Overall accuracy however falls in both cases by 0.02, an insignificant decrease.

It would appear that the application of g-index and h-index is only useful when no citation-context is included within the corpus documents. The improved combinations' rankings are in fact still not as highly correlated to the expert rankings as the baseline TF-IDF rankings within the SIGIR_Comb corpus. Unlike a measure such as PageRank, or TF-IDF, the g-index and h-index do not just take into account the features of a single document (be that links or content). Instead they ignore these features entirely,

[3]Recall that in all tests which follow for statistical significance, we have used a 1-tail paired T-test.



(a) h-index correlation within SIGIR_Txt

(b) g-index correlation within SIGIR_Txt

(c) h-index correlation within SIGIR_Comb

(d) g-index correlation within SIGIR_Comb

Figure 66: Comparisons of the average correlations achieved through combination of baseline TF-IDF scores from the two Lemur indexes with g-index and h-index scores, and the experts' rankings

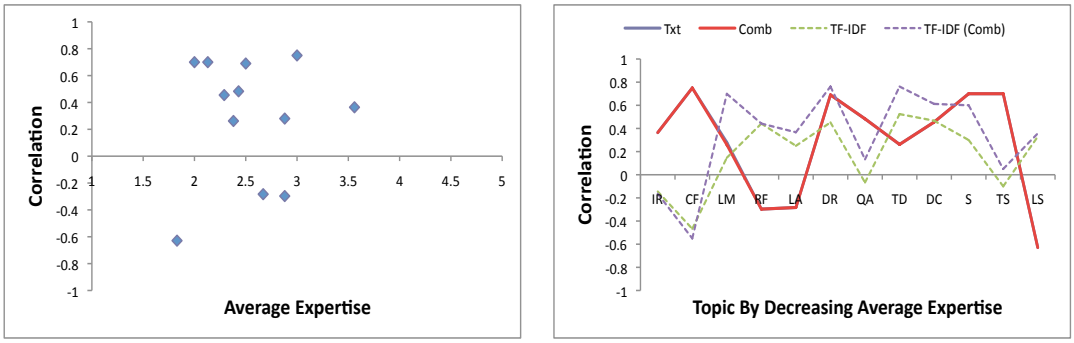
focussing on the links and output of an author. A consequence of this is that more noise may be introduced; authors who are prolific, but not in topics relevant to the field, may be given over-inflated importance. This issue is discussed in the final chapter. Also, as stated previously in Section 4.1.2.1, these particular measures are focussed on *"the number of the papers in the productive core"* of an author (Bornmann et al., 2008). We aim to utilise not just the output, or productivity of an author as an aid in re-ranking papers/comments, but also their impact on the topic of focus. For this reason, we will now look at the m-index also.

Recall that the m-index of an author is aimed at taking into account the impact of those papers/comments from within the productive core, or Hirsch core, of an author. It is defined as:

$$\{h_1, h_2 \dots h_n\} h_i \in H, m = h_{\frac{n}{2}} \tag{31}$$

where H is the ordered list of an author's Hirsch core. i.e. m is the median number of citations received by papers within an author's Hirsch core.

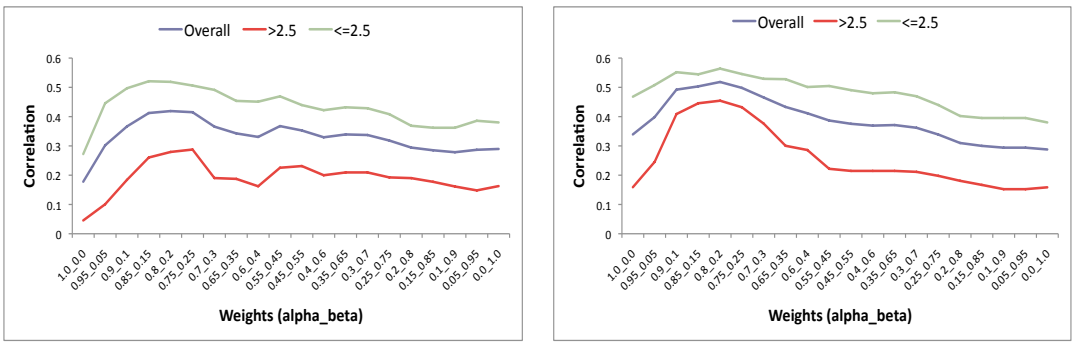
The performance of the m-index in re-ranking the initial set of documents returned



(a) (b)

Figure 67: Comparisons of the m-index re-rankings of the TF-IDF baseline from the two Lemur indexes with the experts' rankings

by the baseline TF-IDF algorithm is markedly better than the previous two author measures. Its improvement on the baseline correlations for the higher expertise rankings is especially noticeable in Figure 68. Its performance is better in nearly all topics, with the exception of *"relevance feedback"* (RF) and *"spam"* (S). The m-index attempts to give credit to authors who are not only highly cited, but also takes into account the median number of citations an author receives. In doing so, it provides a greater differentiation between authors of the same g- or h-index; an author who has written many seminal papers is given more credit than one who has published many well received papers. The fact noticed by Jin (2006) that citation counts can be highly skewed is also considered; an author who has published one seminal work, but never published again is not given as much credit as a more active author. In this way active participation is encouraged. With regards to the SportsAnno or Annoby scenarios, this may be seen as differentiating between an author who has left one or two highly controversial comments, and an author who participates regularly in conversations and community activity.



(a) m-index correlation within SIGIR_Txt (b) m-index correlation within SIGIR_Comb

Figure 68: Comparisons of the average correlations achieved through combination of baseline TF-IDF scores from the two Lemur indexes with m-index scores, and the experts' rankings

Looking at the performance of the m-index in improving correlation with the experts'

ranking, we see that this time the improvement gained from combination with the baseline TF-IDF ranking is considerable though not significant ($p = 0.128$). It is also a gain which is noticeable across both Lemur indexes. Both SIGIR_Txt ($\alpha = 0.75$, $\beta = 0.25$) and SIGIR_Comb ($\alpha = 0.8$, $\beta = 0.2$) benefit from an increase in correlation with the experts' ranking of 0.24 and 0.29 respectively when the TF-IDF score is combined with the respective m-index scores. The cost of this increase in correlation with the experts' ranking is a fall in AP of 0.09 from 0.64 to 0.55; this however is not significant. Unlike the g-index and h-index, the m-index appears to be complementary to the inclusion of citation-contexts also. At worst, the complete re-ranking of the initial TF-IDF ranking within the SIGIR_Comb index by m-index scores decreases the correlation with the experts' ranking by 0.05. This is for overall average, whilst the average correlation of the higher expertise topics remains the same. In the case of the SIGIR_Txt index, the ranking created by m-index alone increases correlation with expert rankings overall by 0.11, and within the higher expertise topics by 0.14. None of these changes however are significant.

Table 19: Optimal combinations achieved for the state-of-the-art citation measures.

Feature	*SIGIR_Txt*				*SIGIR_Comb*			
	α	β	Corr.	A.P.	α	β	Corr.	A.P.
h-index	0.95	0.05	0.08(0.03)	**0.65(0.11)**	1.0	0.0	0.16(-)	0.70(-)
g-index	0.95	0.05	0.03(-0.02)	**0.65(0.09)**	1.0	0.0	0.16(-)	0.70(-)
m-index	0.75	0.25	0.29(0.24)	0.55(0.09)	0.8	0.2	0.45(0.29)	0.55(0.09)

We have shown that, with regards to the extended SIGIR corpus, it would appear the most effective measure currently used to attribute authority to authors (that can then be used in conjunction with standard document weighting TF-IDF to provide a more 'expert' ranking) is the m-index. The m-index attempts to provide a more impact-based measure of an author's work. We hope to achieve this when combining our chosen features with the TF-IDF baseline. We discuss the impact of each of these features in the following section.

6.3 Calculation of Author Feature Contributions

As we have seen, the way in which the impact and authority of an author is measured can have a large effect on the influence the author exerts on the ranking of documents. While g-index and h-index focus on the productive prowess of an author, m-index focusses on the impact that productivity has on an author's surrounding network. We would like to mirror the way in which m-index is more discerning in its valuation of an author, whilst still giving credit to prolific authors. In real-world Web 2.0 scenarios, this means

that we would like to enable users to become noticed for the quality content which they produce, while not giving credit to, effectively, spammers. To do this, we have identified a number of features as put forth in Section 2.4.3.1, proposed originally in the data quality literature by Zhu and Gauch (2000) in the context of web-page quality. We have adapted these features to our own purpose, giving us a group of features as follows:

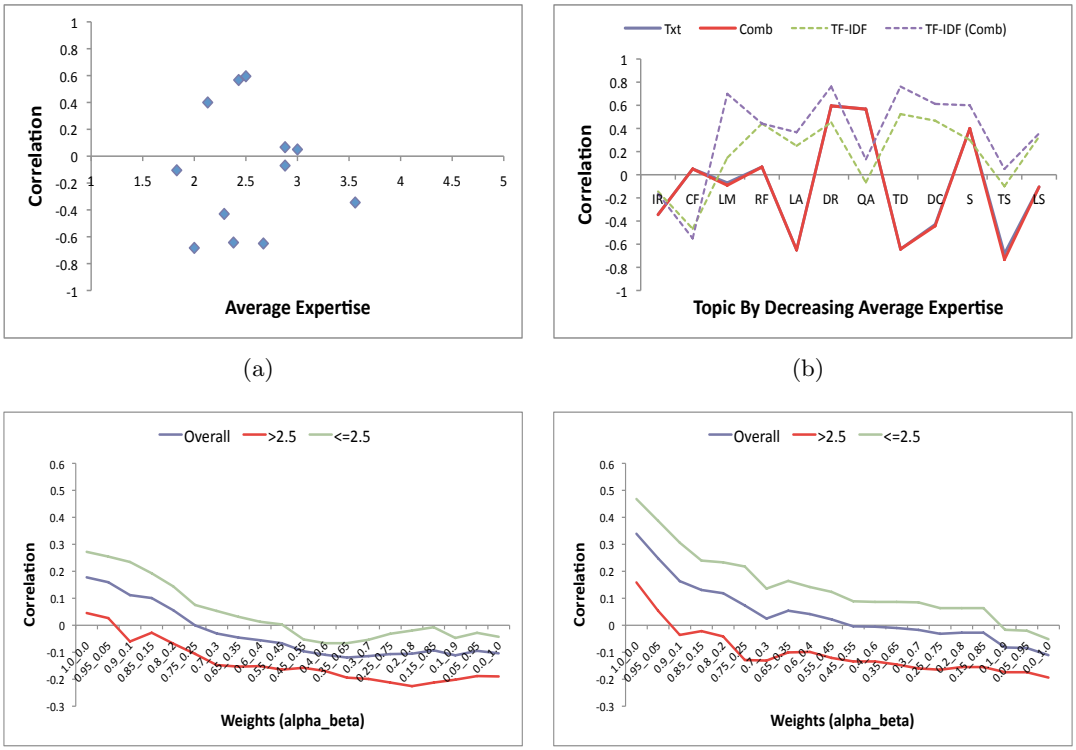
- **Total Comments:** This is the total number of comments created by an author within the corpus. Within the SIGIR context, this refers to both the number of papers written, as well as citations made. In the Web 2.0 context, or indeed blogs, this would typically be the number of comments made, as well as the number of, say, blog posts created. This is further broken down into the following:
  - **Started:** This is the total number of comments created by an author within the corpus which link directly to a source of information. Within SIGIR this means the number of papers an author has written that have not cited a SIGIR paper; within the Web 2.0 scenario this would be either blog posts created, or comments that quote and comment on a source (e.g. blog post, newspaper article etc.) directly. In the case of SportsAnno or Annoby, this refers to any comments whose parent is the original newspaper article, or video.
  - **'Started Threaded':** This refers to the total number of 'started' comments which have received a citation/reply.
  - **'Started Barren':** Those 'started' comments/papers which have gone uncited or without reply.[4]
  - **Replied:** The total number of comments made by an author within the corpus which cite other documents within the corpus. In the SIGIR corpus, this refers to any paper from within the SIGIR corpus which cites another SIGIR paper. Within the Annoby/SportAnno context, this refers to any comment which is in reply to another user's comment, and therefore not directly made on the original content.
  - **'Replied Threaded':** This refers to the total number of 'replied' comments which have received a citation/reply.
  - **'Replied Barren':** Those 'replied' comments/papers which have gone uncited or without reply.

[4]We note that taking the direct value, and not the inverse of the number of barren messages may seem counter-intuitive. It is recognised as a bad sign if a message receives no replies, meaning that there is no interest in the content of the message. We do not take the inverse of this value however, as doing so would effectively promote the notion of publishing sparsely. This is because the more barren messages an author has, the more heavily penalised. We take the approach that while a barren message is not as good as a threaded message, it is still better than not writing a message at all. By taking the count of message within the 'started barren' feature group directly, weighting these accordingly, we give credit to an author who is published highly, but not cited over an author who does not publish.

- **Average Words (log):** This is the average number of words written by an author in a comment. This refers to the citation-context length within the SIGIR context, and more obviously the comment length within the Annoby/SportsAnno contexts. We take the log of this total so as to smooth the effect of the word count.
- **Average Responses (log):** This is the average number of responses received by an author in reply to any comment they might make. This includes any replies received to either 'started' or 'replied' comments, as well as all citations within a thread of paper citations below this author's paper. Again, we take the log of this value to smooth the effect of large number of responses. This is particularly necessary in the SIGIR case, since some older papers have received a large number of citations.

These different features aim to take into account the different aspects of an author's interactions with their social network. Before being able to combine them however, we must first examine the effect which each of these individual features has on the correlation of our baseline TF-IDF ranking. The first of these is the total number of comments as shown in Figure 69. We can see that there is no real correlation between the expertise and expert ranking correlation. This fact is exemplified in the effect of combination of the total number of comments with the TF-IDF baseline score. The 'total comments' feature alone provides almost no correlation with the experts' ranking, and combination with the TF-IDF score is universally insignificant, but detrimental within both the SIGIR_Comb and SIGIR_Txt indexes, over all correlations as seen in Figure 69(c) and 69(d). AP is significantly affected ($p = 0.043$) by re-ranking the result set based purely on 'total comments', falling from 0.64 to 0.17.

Next we look at the effect of average words per comment/citation on the ranking correlation. Re-ranking the set of documents based on the average citation-context length of an author creates a ranking which is in fact negatively correlated with the average expertise of the expert rankings. This may be seen in Figure 70(a). Figure 70(b), which shows that the effect of re-ranking solely based on the log average word count is negative or near-zero correlation in all of the top-expertise topics. The effect of inclusion of the 'average words' feature is universally detrimental on the SIGIR_Comb index. Though it provides a very small boost in the SIGIR_Txt index by setting $\alpha = 0.85$ and $\beta = 0.15$, this improvement of 0.04 is not significant ($p = 0.331$). The effect of words within our SIGIR corpus may be dampened by the fact that the citation-contexts were chosen heuristically to be three sentences in length. Although this does provide some variety, since it only splits on full-stops, there is still far less variation than in a real-world situation. For this reason we advise caution in completely disregarding the average words written by an author as an indicator of author's worth. In our case however, it does not seem to prove effective.



(a)

(b)

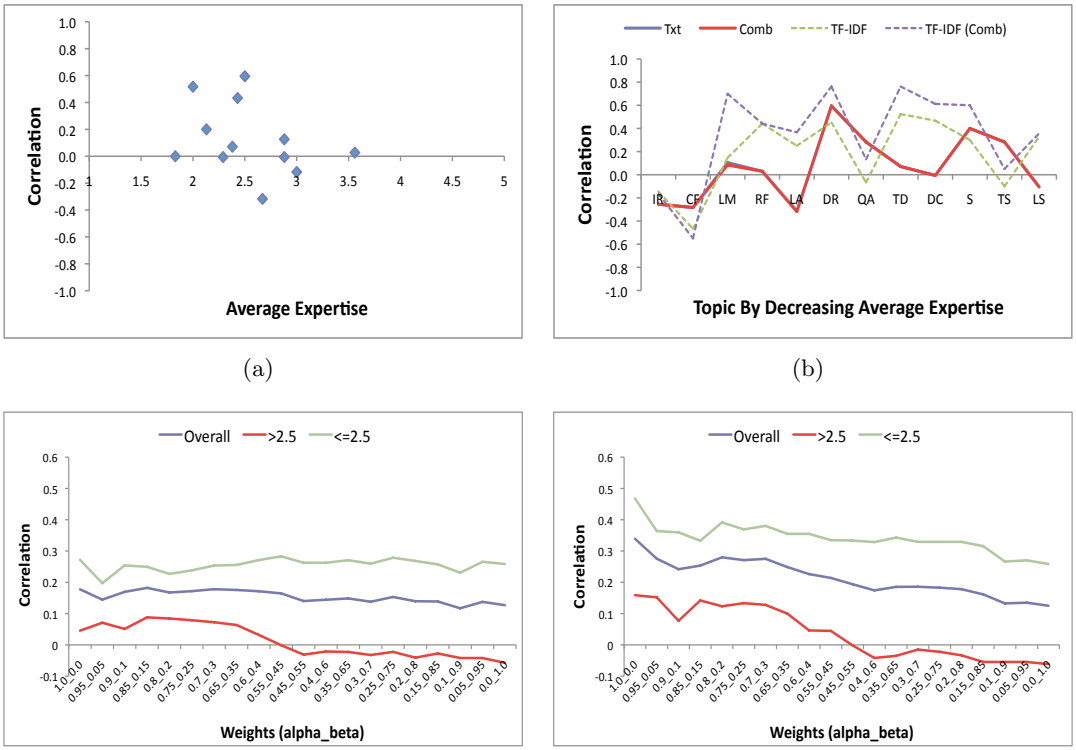
(c) Experts' ranking correlation within the SIGIR_Txt index

(d) Experts' ranking correlation within the SIGIR_Comb index

Figure 69: Comparisons of the 'total comments' re-rankings of the TF-IDF baseline from the two Lemur indexes with the experts' rankings.

We now look at the effects of splitting the comments created by an author/user into *started* and *replied*, and considering each of these in turn. These two groupings are split once more into *threaded* and *barren* subgroups which we also consider. First the initial split of started and replied which may be seen in Figure 71.

We can see from Figure 71(a) that there is a slight negative correlation between the average expertise of the experts' ranking, and the rankings provided by the 'started' feature. We see that any inclusion of the 'started' feature in the SIGIR_Comb index is detrimental to performance. Interestingly, setting $\alpha = 0.9$ and $\beta = 0.1$ yields an improvement in correlation for the SIGIR_Txt index. This improvement of 0.08 however is not significant ($p = 0.152$), being made in the lower expertise topic correlations, and bringing the rankings more in line with those of a novice rater. This is not something we want, so again we see that setting $\alpha = 1.0$ and $\beta = 0.0$ provides the best correlation. The increase in lower expertise correlation also significantly ($p = 0.012$) reduces the AP for lower expertise topics by 0.17, from 0.72 to 0.55. One possible explanation for the poor performance of this feature is that in the context of the extended SIGIR corpus, 'started' comments refer to the authoring of a paper which is in the SIGIR corpus, but does not cite any other SIGIR papers, as doing so would place that paper within the 'replied' subgroup. Since authors often cite papers from within the proceedings of a conference that they wish to have a paper accepted for, there are subsequently fewer



(a)

(b)

(c) Experts' ranking correlation within the SIGIR_Txt index

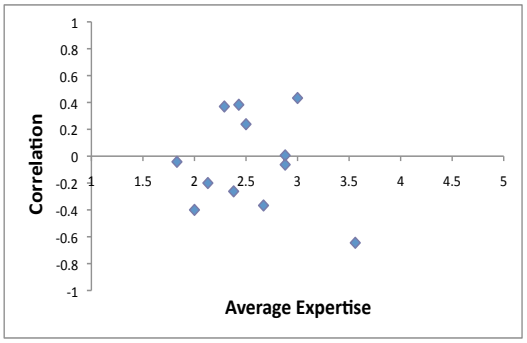
(d) Experts' ranking correlation within the SIGIR_Comb index

Figure 70: Comparisons of 'log average words' re-rankings of the TF-IDF baseline from the two Lemur indexes with the experts' rankings.

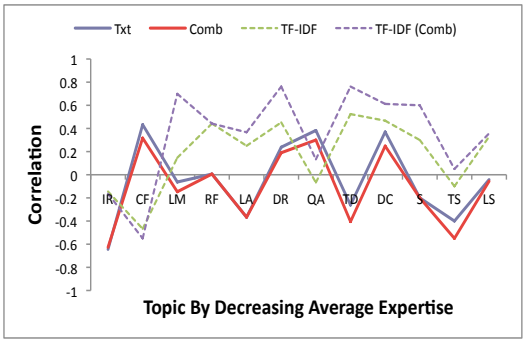
papers that will appear within this 'started' subgroup than the complementary 'replied' subgroup. Again, this may be seen as a weakness of the corpus itself. The 'started' feature does provide a benefit to the ranking correlation of lower expertise topics within both indexes, as shown in Figures 72(a) and 72(c). The 'replied' feature in Figure 71(c) does not however show any significant correlation with average expertise.

Since a large number of papers go without ever being cited (as shown in Figure 73), it is useful to look at the contributions of both the threaded, or cited, papers as well as the barren un-cited ones separately. In a real-world scenario this distinction should also be made. An author/user who writes large numbers of comments that are largely ignored is not of any great importance. The comments that they write may however simply be the last comment in a thread which again will receive no replies, but are part of a larger conversation. Here we consider the subgroups 'threaded' and 'barren' from within both previous groups 'started' and 'replied' to see whether these subgroups can better help to distinguish between authors' contribution and impact.

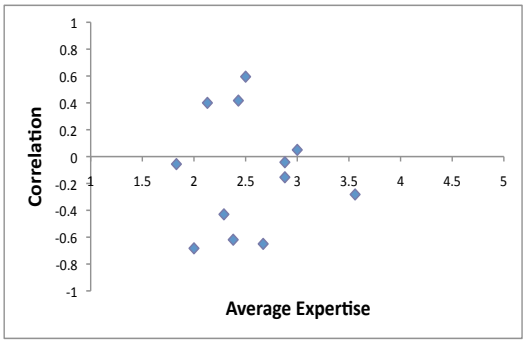
The correlations shown by re-ranking the initial returned set from a TF-IDF query, solely based on the scores obtained from the 'started' subgroup 'barren' are positively correlated with the average expertise as shown in Figure 74(a). This subgroup consists of those papers which have not referenced another SIGIR paper, and have never been



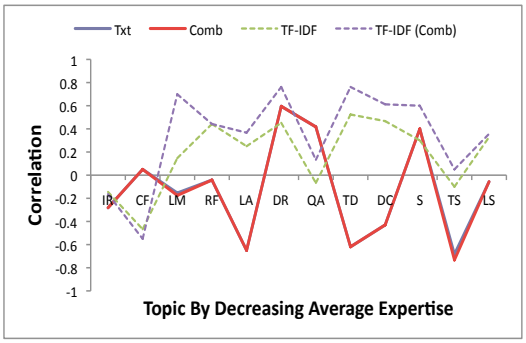
(a) 'started' correlation with experts' ranking



(b) Correlation of 'started' re-ranking with experts' ranking



(c) 'replied' correlation with experts' ranking

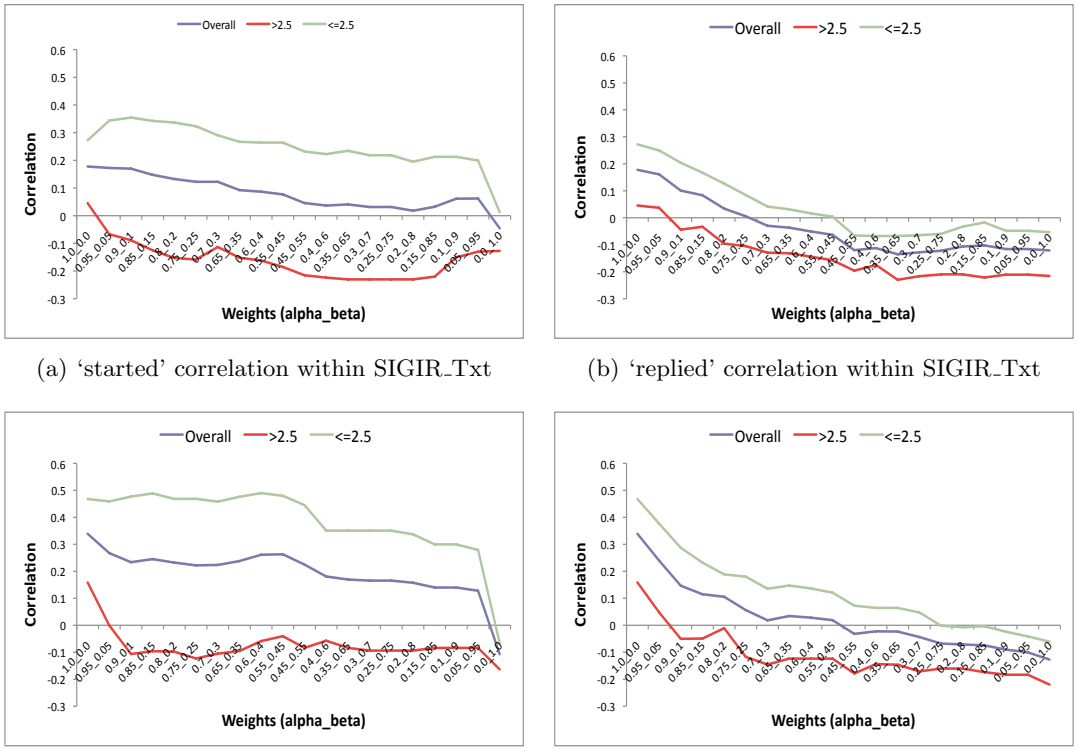


(d) Correlation of 'replied' re-ranking with experts' ranking

Figure 71: Comparisons of the 'started' and 'replied' re-rankings of the TF-IDF baseline from the two Lemur indexes with the experts' rankings

referenced themselves. Including these in the calculation of an authors' impact or importance dampens the rankings of authors who have many papers which have not been referenced. They do not 'fit' into the proceedings since they do *not* reference any other SIGIR papers. Figure 74(c) shows that correlation within the higher expertise topics on the SIGIR_Txt is improved by 0.05 when setting $\alpha = 0.75$ and $\beta = 0.25$, but this is not significant ($p = 0.327$). The corresponding settings reduce AP significantly however ($p = 0.021$), from 0.72 to 0.49. Again, combination of this feature on the SIGIR_Comb index results in detrimental performance.

Focussing now instead on the 'threaded' subgroup, we consider papers from within the SIGIR corpus, citing no SIGIR papers, but receiving citations themselves. Since we have chosen a fixed time-frame from within the SIGIR proceedings, there are a relatively small number of these papers. Most of these papers come from the earier years (1997, 1998) and so can only be credited to authors who have been publishing in SIGIR for a significant portion of our time-frame. This time we see that any inclusion of the 'threaded' feature in either index is detrimental to performance (Figures 74(g) and 74(h)). This may seem strange (effectively removing any benefit from the creation of new papers which do not cite past SIGIR publications), but from the standpoint of a Web 2.0/social media scenario it is not harmful. We are effectively penalising those



(a) 'started' correlation within SIGIR_Txt

(b) 'replied' correlation within SIGIR_Txt

(c) 'started' correlation within SIGIR_Comb

(d) 'replied' correlation within SIGIR_Comb

Figure 72: Comparisons of the average correlations achieved through combination of baseline TF-IDF scores from the two Lemur indexes with the 'started' and 'replied' scores, and the experts' rankings

authors who do not reference or engage in any way with the community. This is similar to a user who leaves comments showing their own opinion, but will never comment on or answer their critics.

Looking at the correlation scores created by using just the 'replied' subgroups' re-ranking of the initial TF-IDF set, shown in Figures 75(a) and 75(e), we see a different picture to those of Figures 74(a) and 74(e). The inclusion of 'barren' feature information is detrimental to the correlation scores, and consequently, correlation scores are best when $\alpha = 1.0$ and $\beta = 0.0$. This could be due to the fact that even though these papers from within the 'replied' subgroup are part of a larger citation thread of papers, they would not be included within the g-, h-, or m-index of the author; they are effectively noise. The slight boost that they can provide within the lower expertise topics of the SIGIR_Txt index may be viewed in a number of ways; one way is to consider a person who cites many people but is not as highly cited themselves. They may be seen as a key into an important group of authors, whilst not necessarily being part of the grouping themselves.

The performance of the 'threaded' subgroup is far better. We can see that nearly all of the correlation scores created by re-ranking the initial TF-IDF set are higher than those of the 'barren' feature. This set consists of papers which have cited other SIGIR

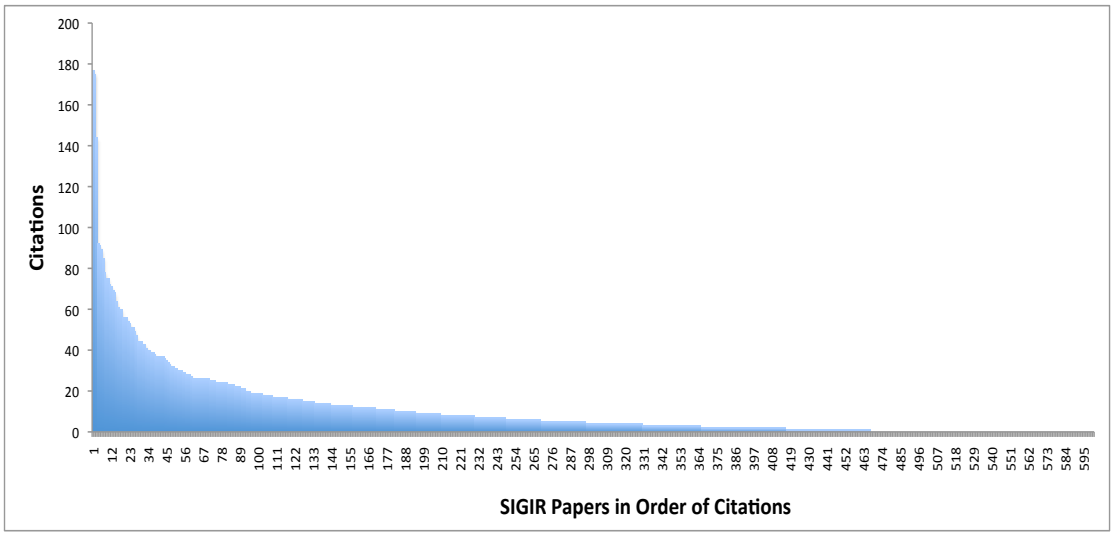


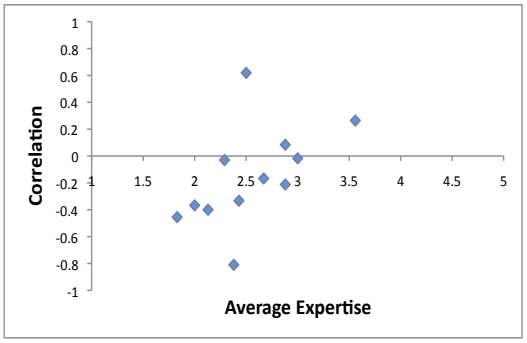
Figure 73: The number of citations received by each of the SIGIR papers in our extended corpus. We can see that there are many papers with few to no citations.

papers, whilst they themselves are cited. As such, they draw on a history of SIGIR research and have been themselves cited as useful research [5]. The other important difference between the 'barren' and 'threaded' subgroups is that papers which are barren do not benefit at all from the techniques used to create the SIGIR_Comb index; since they have no citations, there is no citation-context to add to the document, and therefore no additional index terms.

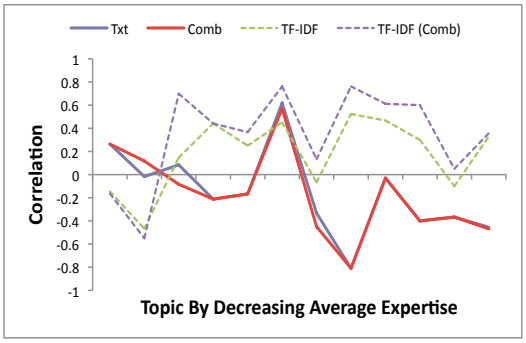
Looking at the correlation scores for combinations of the 'replied' subgroup 'threaded' we see that, for the SIGIR_Comb index, a significant improvement ($p = 0.071$) can be made by setting $\alpha = 0.95$ and $\beta = 0.05$. The increase attained in AP by using these weights is not significant ($p = 0.248$). This is repeated in the SIGIR_Txt index, where the inclusion of 'threaded' feature information has a large positive effect on the correlation of rankings, increasing the overall correlation between the returned ranking and the experts' ranking by a maximum of 0.06, whilst increasing the higher expertise topics' correlation by 0.08. This is achieved by setting $\alpha = 0.7$ and $\beta = 0.3$. This increase in correlation is not significant ($p = 0.198$), neither is the increase in AP from 0.54 to 0.57 ($p = 0.345$). It would appear however that the effect of the 'replied threaded' feature is not apparent when citation-context is included. This may be because the benefit of the threading is already shown in a more direct manner than through the author; any threading of a paper through citation by other papers provides that paper with additional index terms in its own right. These citations then boost the underlying TF-IDF baseline, masking the effects of the 'replied threaded' feature.

The last feature we shall look at is the average number of responses an author receives to comments/citations made. In the SIGIR corpus, this translates to the average number

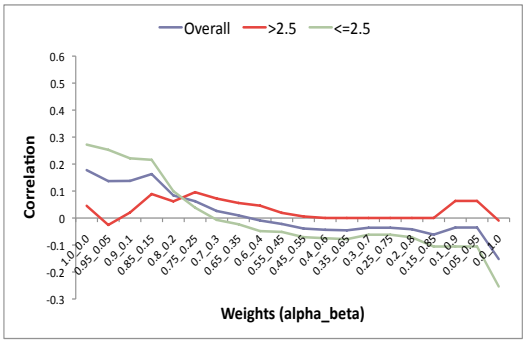
[5] It may not be the case that the citation is for a positive reason, but we argue that this is also useful. Future research is informed not just by the successes of the past, but also by the failures.



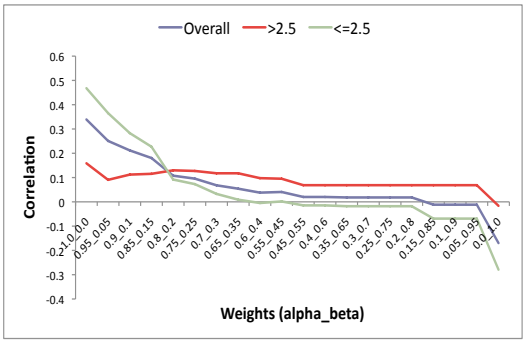
(a) 'barren' correlation with experts' ranking



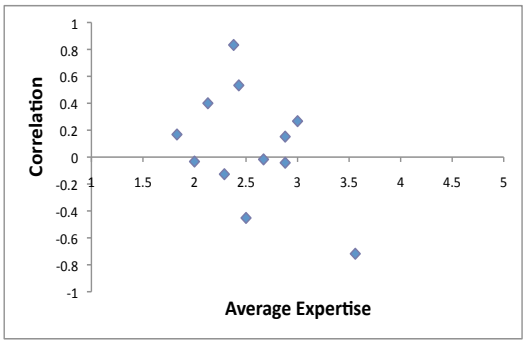
(b) Correlation of 'barren' re-ranking with experts' ranking



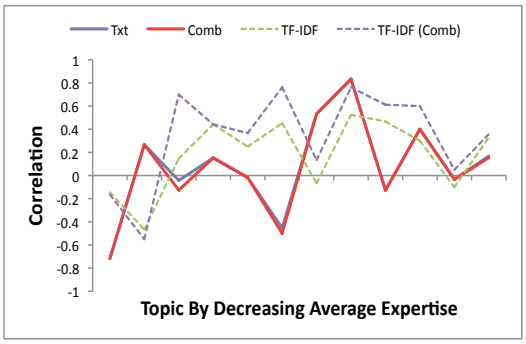
(c) 'barren' correlation within SIGIR_Txt



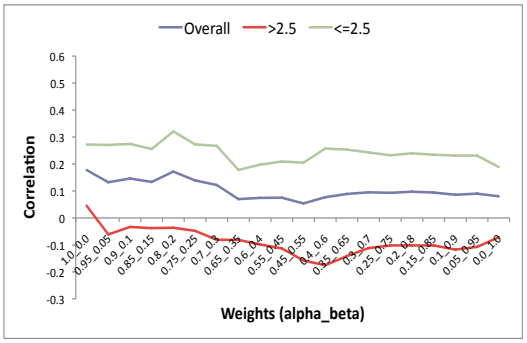
(d) 'barren' correlation within SIGIR_Comb



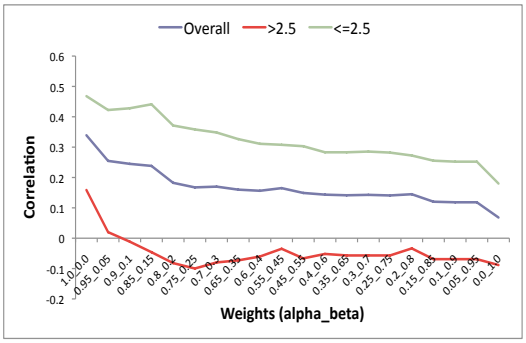
(e) 'threaded' correlation with experts' ranking



(f) Correlation of 'threaded' re-ranking with experts' ranking

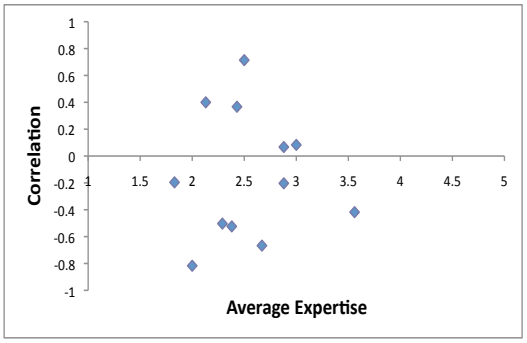


(g) 'threaded' correlation within SIGIR_Txt

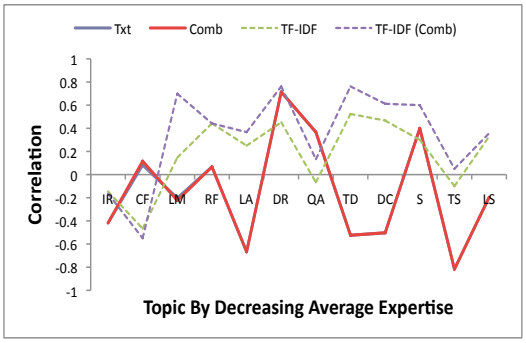


(h) 'threaded' correlation within SIGIR_Comb

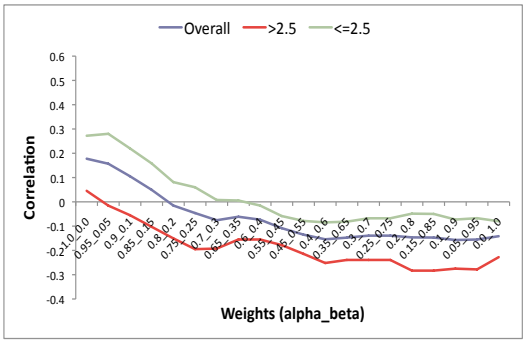
Figure 74: Comparisons of the average correlations achieved through combination of baseline TF-IDF scores from the two Lemur indexes with the 'started' subgroups 'barren' and 'threaded' scores, and the experts' rankings



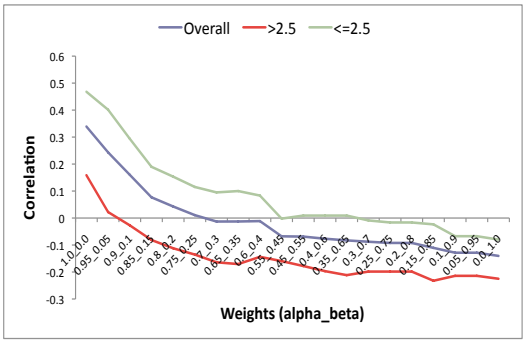
(a) 'barren' correlation with experts' ranking



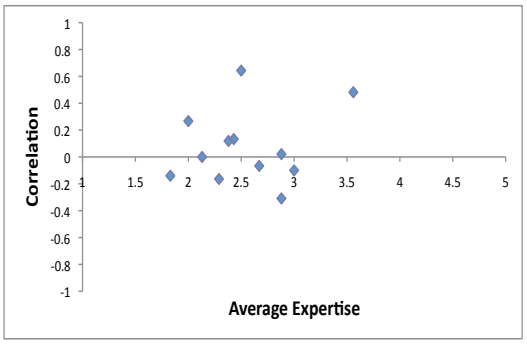
(b) Correlation of 'barren' re-ranking with experts' ranking



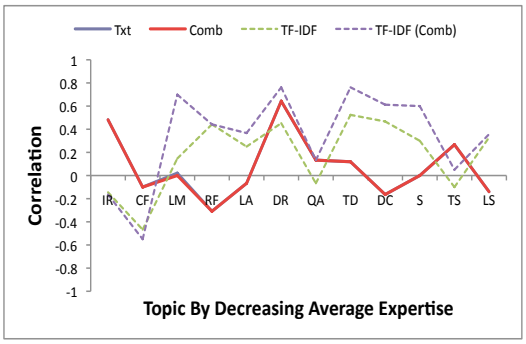
(c) 'barren' correlation within SIGIR_Txt



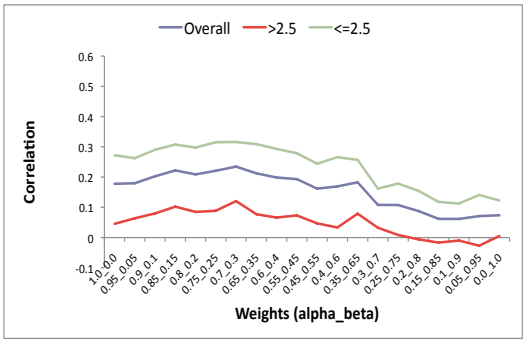
(d) 'barren' correlation within SIGIR_Comb



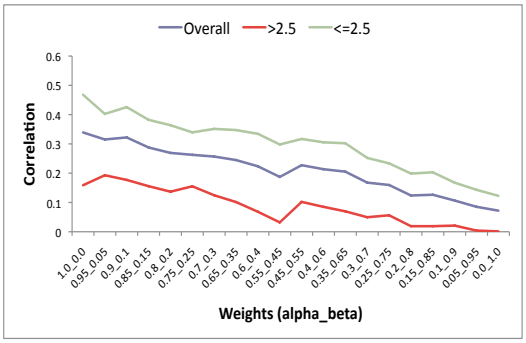
(e) 'threaded' correlation with experts' ranking



(f) Correlation of 'threaded' re-ranking with experts' ranking

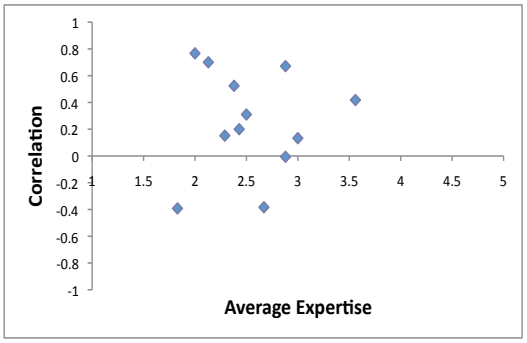


(g) 'threaded' correlation within SIGIR_Txt

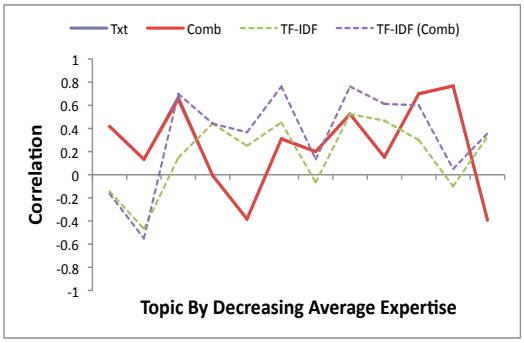


(h) 'threaded' correlation within SIGIR_Comb

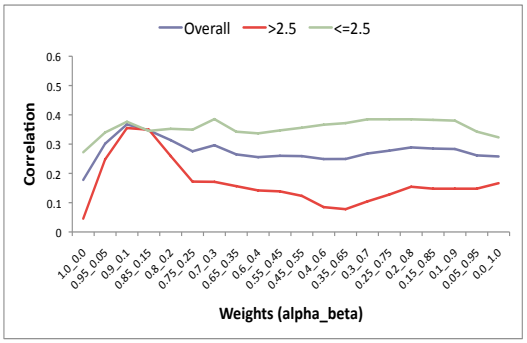
Figure 75: Comparisons of the average correlations achieved through combination of baseline TF-IDF scores from the two Lemur indexes with the 'replied' subgroups 'barren' and 'threaded' scores, and the experts' rankings



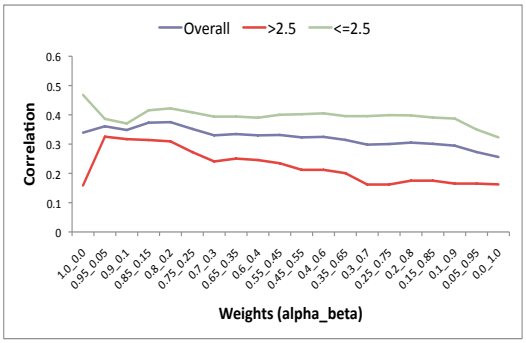
(a) 'responses' correlation with experts' ranking



(b) Correlation of 'responses' re-ranking with experts' ranking



(c) Experts' ranking correlation within the SIGIR_Txt index



(d) Experts' ranking correlation within the SIGIR_Comb index

Figure 76: Comparisons of the log 'average responses' re-rankings of the TF-IDF baseline from the two Lemur indexes with the experts' rankings.

of citations received by each of the author's papers. This is similar to the m-index, but does not discriminate against papers which are not contained within the Hirsch core. We can see from Figure 76(a) that there is a slight positive correlation between average expertise of experts' rankings and the rankings created by the log 'average responses' feature. The inclusion of log 'average response' feature information in the re-ranking procedure is beneficial to both SIGIR indexes. Setting $\alpha = 0.9$ and $\beta = 0.1$ provides a significant improvement ($p = 0.026$) to the baseline TF-IDF rankings in the SIGIR_Txt index. Values for $\alpha = 0.95$ and $\beta = 0.05$ again significantly improve the correlations within the SIGIR_Comb index ($p = 0.074$). From this we can ascertain that the inclusion of an author's 'log average response' information is as effective as citation-contexts in raising the correlations of between rankings more expert users' rankings, and that returned by an IR system (as shown in Figures 76(c) and 76(d)). Using the new combination, AP for SIGIR_Txt remains fixed at 0.54, however there is a decrease in SIGIR_Comb from 0.61 to 0.58 but it is not significant ($p = 0.3$).

A count of average responses, similar to threaded comments, gives an idea of the popularity of an author within their social network. This is akin to PageRank, which considers each link to a page to be a vote for that page. Citations are similar in that citing a paper infers some kind of influence or impact of that paper on the current

work. Within the context of the SIGIR_Txt index we can see that those features which measure interaction with the community, or are of a link-based nature (e.g. threaded replies and average responses), perform best in creating a more expert-like ranking. Indeed log 'average responses' is the only measure which improves correlation on both indexes significantly. It should be remembered however, that unlike PageRank, the measures are taken on the authors and not on any specific document.

Table 20: Optimal combinations achieved for single author feature measures.

Feature	*SIGIR_Txt*				*SIGIR_Comb*			
	α	β	Corr.	A.P.	α	β	Corr.	A.P.
Comments	1.00	0.00	0.05(-)	0.65(-)	1.00	0.00	0.16(-)	0.70(-)
Avg. Words	0.85	0.15	0.09(0.04)	**0.57(-0.08)**	1.00	0.00	0.16(-)	0.70(-)
Started	1.00	0.00	0.05(-)	0.65(-)	1.00	0.00	0.16(-)	0.70(-)
Threaded	1.00	0.00	0.05(-)	0.65(-)	1.00	0.00	0.16(-)	0.70(-)
Barren	0.75	0.25	0.10(0.05)	**0.49(-0.23)**	1.00	0.00	0.16(-)	0.70(-)
Replied	1.00	0.00	0.05(-)	0.65(-)	1.00	0.00	0.16(-)	0.70(-)
Threaded	0.70	0.30	0.12(0.07)	0.57(0.03)	0.95	0.05	**0.19(0.03)**	0.63(0.02)
Barren	1.00	0.00	0.05(-)	0.65(-)	1.00	0.00	0.16(-)	0.70(-)
Avg. Responses	0.90	0.10	**0.36(0.31)**	0.54(0.00)	0.95	0.05	**0.33(0.17)**	0.58(-0.03)

6.3.1 Combination of Author Features

Now that we have discovered the impact of each of the individual features, we would like to find some way to combine these features. The first and easiest way to do this is by using the same technique as in the previous section; we perform a weighted linear combination of the features discussed above. We do not however include all the features, since doing so would give additional influence to certain feature groups. We include the subgroups of replied and started messages, but not the actual features themselves. Our final weighted combination is made up of all the features discussed above, minus the 'total comments', 'started' and 'replied' features. This is because each of the included features belongs to a subgroup of the excluded features. Removing these three features leaves us with six features to be combined in such a way as to provide an optimal correlation with the experts' ranking.

Table 21 shows the optimal combinations of weights for each of the topics. These are the features of the author of a paper, taking into account the number of papers an author has written, the number of citations an author has received, and the amount an author has written on average as 'citation-context'. It may be seen that the log(words) feature plays almost no role in any of the optimal weight combinations, with the exception of the Distributed Retrieval (DR) topic. This is the average number of words written in a citation-context by an author, something which none of the author metrics in Section

Table 21: Optimal per-topic weights for linear combinations of the author features.

(a) SIGIR_Txt weights

Topic	*Weights*						
	log(W)	SB	ST	RB	RT	log(R)	Corr.
IR	0.05				0.60	0.35	0.827
CF	0.05	0.40	0.30	0.05	0.05	0.15	0.633
LM	0.20			0.25		0.55	0.804
RF	0.05	0.20	0.70			0.05	0.588
LA	0.10				0.80	0.10	0.067
DR	0.80					0.20	0.905
QA		0.20	0.35	0.30		0.15	1.000
TD			0.65	0.20	0.10	0.05	0.881
DC		0.40	0.35	0.20	0.05		0.527
S	0.35	0.10	0.15	0.20	0.15	0.05	1.000
TS			0.20	0.10	0.55	0.15	0.952
LS	0.10	0.10	0.35	0.05	0.40		0.552

(b) SIGIR_Comb weights

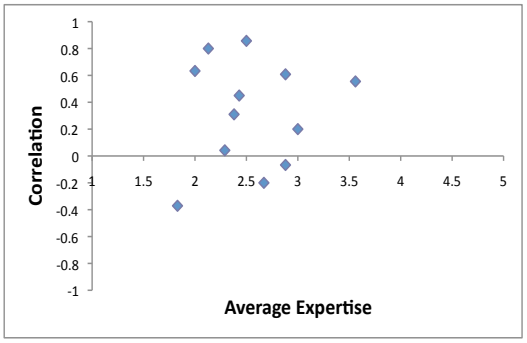
Topic	*Weights*						
	log(W)	SB	ST	RB	RT	log(R)	Corr.
IR	0.05				0.60	0.35	0.827
CF	0.05	0.40	0.30	0.05	0.05	0.15	0.633
LM	0.20	0.05		0.20		0.55	0.783
RF	0.05	0.20	0.70			0.05	0.588
LA	0.10				0.80	0.10	0.067
DR		0.45	0.20		0.10	0.25	0.929
QA		0.20	0.35	0.30		0.15	1.000
TD			0.65	0.20	0.10	0.05	0.881
DC		0.40	0.35	0.20	0.05		0.527
S			0.20	0.15	0.10	0.55	1.000
TS	0.15	0.05	0.30			0.50	0.867
LS	0.25		0.40	0.15	0.05	0.15	0.391

6.2 take into account. Indeed, inclusion of the citation-contexts within each document's bag-of-words as in SIGIR_Comb, removes the influence of words completely.

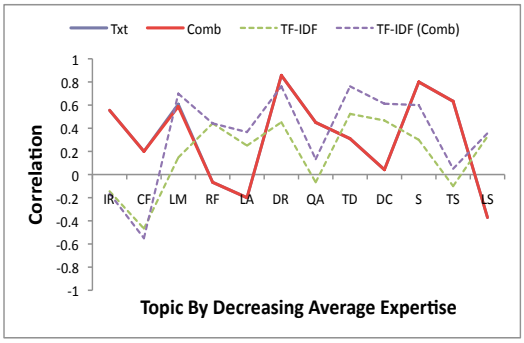
Neither of the two 'barren' subgroups play a significant role in the higher expertise topics, except in the cases of the 'collaborative filtering' (CF) and 'document clustering' (DC) topics. In these two topics, they also receive higher weights than those of the 'threaded' subgroups. One reason for this may be the referencing of seminal papers by several papers which are themselves not well cited; this could be due to a narrowing of the field on one particular point during our window of time. In the case of collaborative

filtering, another explanation could be the abundance of conferences which contain research on this topic. Collaborative filtering is less specific to SIGIR and so many papers which are cited by papers in this area will not be from the SIGIR proceedings. This makes it more likely that the highly ranked papers may be barren, having no citations from SIGIR publications. In the case of document clustering, much of the seminal work in the field happened before the time-window which we are studying.

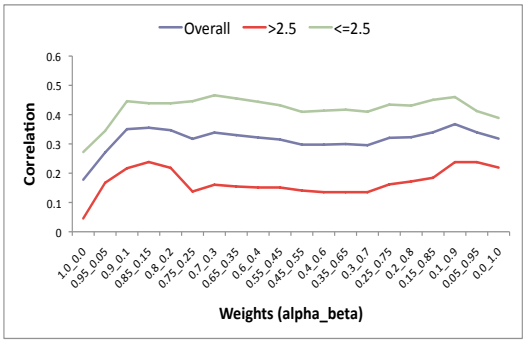
On a topic specific level it would also appear that the number of responses/citations which an author receives is significant in the case of all higher expertise topics, within both indexes with the exception of 'relevance feedback' (RF). This topic however places most weight on the 'started threaded' feature which takes into account the number of papers which an author has written that have subsequently been cited. This pattern is in fact repeated with both the 'link analysis' and 'image retrieval' (IR) topics, except that in these cases the weight is for the 'replied threaded' feature. This may be due to the fact that the most heavily cited paper in these topics are papers found in the later part of the time-window of our corpus. 'Language modelling' (LM) places the greatest importance on the 'responses' feature, not surprisingly since, as we have mentioned in the past, this topic contains the most highly cited document within the corpus; *'A Language Modelling Approach to Information Retrieval'*.



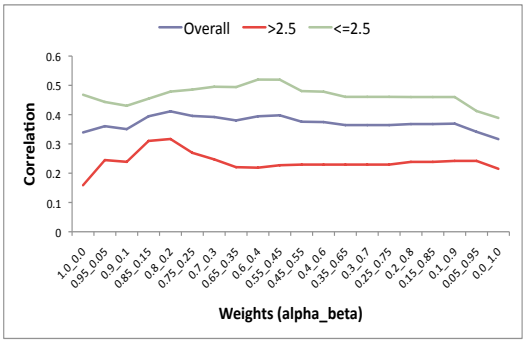
(a) 'aLinear' correlation with experts' ranking



(b) Correlation of 'aLinear' re-ranking with experts' ranking



(c) Experts' ranking correlation within the SIGIR_Txt index



(d) Experts' ranking correlation within the SIGIR_Comb index

Figure 77: Comparisons of the 'aLinear' re-rankings of the TF-IDF baseline from the two Lemur indexes, with the experts' rankings.

While the weights in Table 21 shows the optimal combinations for each individual topic, these results are too specific to a single topic or query. In order to find a combination of weights which is more generally applicable, we must look for those weights which maximise the correlations of all the higher expertise topics' rankings with those of our experts'. Table 22 shows the weights which maximise the most expertise topics, as well as those that maximise all topics respectively. The higher expertise topics place more weight on the number of responses a paper has received. In the case of the SIGIR corpus, this means papers that have been highly cited by highly cited papers. Neither the overall, nor higher expertise topic weighting place emphasis on the 'started' features subgroup. This may well be a consequence of the nature of our data-set.

Table 22: Optimal weights for linear combination of author features across topics

Topics	*Weights*					
	log(W)	SB	ST	RB	RT	log(R)
Top	0.15	0.05		0.05	0.15	0.60
All	0.10			0.20	0.35	0.35

The ability of the author feature linear combination to improve the rankings of higher expertise topics is shown in Figure 77. Setting $\alpha = 0.85$ and $\beta = 0.15$ during retrieval against the SIGIR_Txt corpus results in an increase of correlation with the experts' ranking from 0.04 to 0.24. This increase is insignificant ($p = 0.157$), as is the decrease in AP from 0.65 to 0.62 ($p = 0.326$). Within the SIGIR_Comb index the linear combination produces a slightly smaller increase from 0.16 to 0.32 ($p = 0.181$). This increase is achieved by setting $\alpha = 0.8$ and $\beta = 0.2$, decreasing AP significant from 0.70 to 0.57 ($p = 0.091$).

From the results we have presented, we can see that not surprisingly those papers which have received more responses (thereby becoming threaded) are of greatest importance. This can be seen in the high weighting of the 'average responses' and 'threaded' features. As we have said and will discuss again in Chapter 7, the higher weights given to the 'replied' features as opposed to those of the 'started' features may well be a consequence of our corpus.

6.4 Calculation of AuthorRank Weights

Finding the optimal weights for a linear combination of the features allows us to create a baseline against which to compare our more elaborate combination of features. Linear combination of the values is the simplest way in which the different features may be combined. Now that we have found this, we would like to look at how effective our AuthorRank algorithm is in comparison. AuthorRank attempts to use the features to

give credit to those authors who are highly active within the community, publishing frequently will not being ignored. In effect, it attempts to find a measure of centrality for the author in question; an author who comments/publishes often, and is highly cited/answered is considered to be centrally located within the network. Recall that AuthorRank is of the following form:

$$A_R = \log(Avg_{wc}) * \{\frac{S_T + \alpha * S_B}{S_{TOT}} + \beta * [\frac{R_T + \gamma * R_B}{R_{TOT}}]\} + \log(Avg_r) * [\sum_{x=1}^{n} \frac{r_x}{e^x}] \quad (32)$$

Table 23: Optimal per topic weights for parameter values within the AuthorRank algorithm.

(a) SIGIR_Txt weights

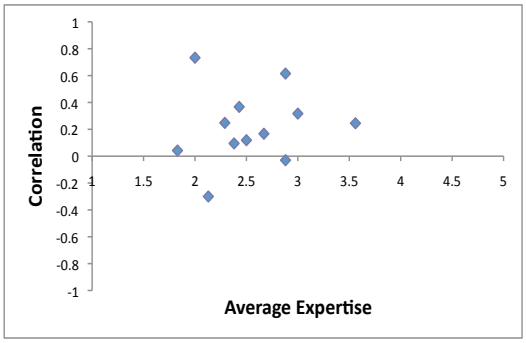
Topic	*Weights*			Corr.
	α	β	γ	
IR	1.00			0.673
CF	0.85	0.15		0.500
LM	1.00			0.769
RF	0.10	0.05	0.85	0.357
LA	0.35	0.60	0.05	0.190
DR	0.65	0.05	0.30	0.833
QA		0.65	0.35	0.483
TD			1.00	0.762
DC	0.55		0.45	0.067
S	0.20	0.75	0.05	0.800
TS	0.15	0.25	0.60	0.857
LS	0.80	0.05	0.15	-0.018
All	0.10	0.05	0.85	
Top	0.10	0.05	0.85	

(b) SIGIR_Comb weights

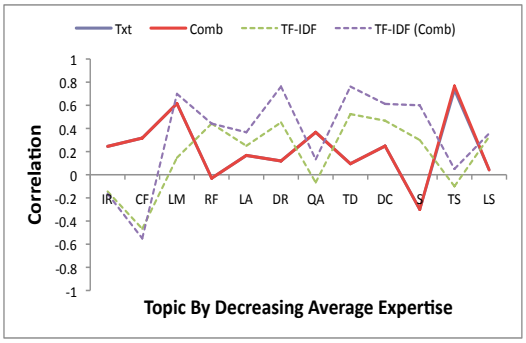
Topic	*Weights*			Corr.
	α	β	γ	
IR	1.00			0.673
CF	0.85	0.15		0.619
LM	0.90		0.10	0.748
RF			1.00	0.067
LA	0.10	0.90		0.033
DR	0.95		0.05	0.857
QA		0.65	0.35	0.483
TD			1.00	0.762
DC	0.20	0.40	0.40	0.238
S		0.05	0.95	0.200
TS	0.35	0.60	0.05	0.857
LS	0.25	0.15	0.60	-0.082
All	0.10	0.20	0.70	
Top	0.10	0.20	0.70	

We can now see that each of the features we have been examining singly is used in combination to make the AuthorRank equation; average word count (Avg_{wc}); average responses (Avg_r); started barren (S_B) and threaded (S_T); and finally replied barren (R_B) and threaded (R_T). In addition to this, AuthorRank also takes into account the fraction of responses which occur at each nested level x below this author's comments/publications. e^x is used as a decay function to limit the effect of responses on the author as they become more highly nested. Using this combination method for the features, we need only find the values for α, β and γ which maximise the correlation of higher expertise topic rankings with those of the experts'. A grid-search allows us to set these parameter as displayed in Table 23.

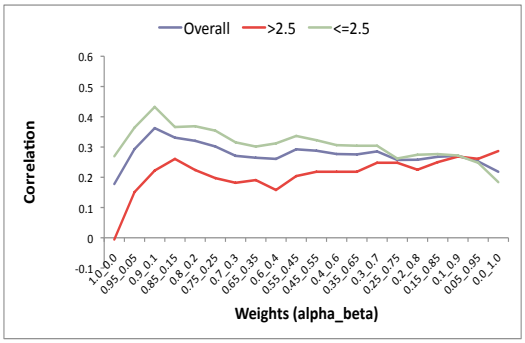
Setting $\beta = 0.0$ means that all information about an author's replies is removed, leaving only the information on the number of threads which an author has begun.



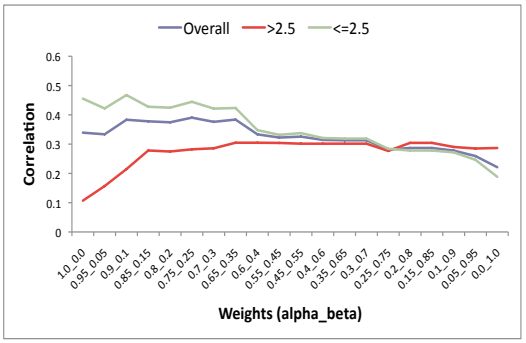
(a) 'aRank' correlation with experts' ranking



(b) Correlation of 'aRank' re-ranking with experts' ranking



(c) Experts' ranking correlation within the SIGIR_Txt index



(d) Experts' ranking correlation within the SIGIR_Comb index

Figure 78: Comparisons of the 'aRank' re-rankings of the TF-IDF baseline from the two Lemur indexes, with the experts' rankings.

This also automatically sets γ to zero, as we can see from Equation (32). The results in Table 23 reveal that the feature 'replies barren' (R_B) plays no part in the optimal weightings for any of the top expertise topics within SIGIR_Comb. Despite this, the optimal combination of weights across all the top expertise topics combined shows that setting $\alpha = 0.10$, $\beta = 0.20$, and $\gamma = 0.7$ obtains the highest correlation between these topic's re-ranked lists and those of the experts' ranked lists.

Using AuthorRank alone to re-rank the TF-IDF baseline results in an improvement in correlation with the experts' ground-truth which is better than the linear combination of all features from the last section. Setting $\alpha = 0.0$ and $\beta = 1.0$, thereby ignoring the influence of TF-IDF completely results in a significant increase ($p = 0.081$) in correlation with the experts' rankings from -0.02 to 0.38 within the SIGIR_Txt index. A significant increase ($p = 0.082$) is also seen in the SIGIR_Comb index when $\alpha = 0.6$ and $\beta = 0.4$, increasing correlation from 0.10 to 0.38. Both indexes do however experience a significant drop in AP from 0.56 to 0.17 ($p = 0.009$), and 0.70 to 0.19 ($p = 0.021$) respectively.

Despite the significant increases with both indexes, there is no significant difference between the improvements in correlation offered by AuthorRan, and that of the straight linear combination of the last section on either index. Figure 79 shows the correlations

per topic for each of the combination techniques. We also include the graph of the best performing single feature 'average responses' as well as the TF-IDF baseline. We can see that AuthorRank differentiates between the depths at which responses are found, penalising them more heavily as they move away from the original author, resulting in poor performance on the 'language modelling' (LM) topic in comparison to the other measures on the SIGIR_Comb index. It does however perform best on the higher expertise topics in general, though as we have stated, this improvement in correlation is not significantly better than that of the linear combination 'aLinear'. With the exception of 'link analysis' (LA) on the SIGIR_Txt topic however, AuthorRank does provide a positive improvement in correlation with the experts' ground-truth.

6.5 The Contribution of Single Messages

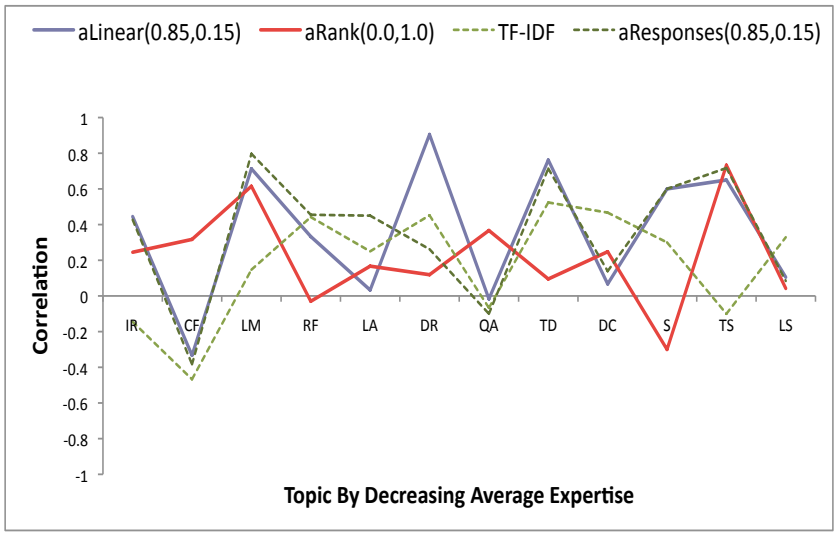
In the above section we have examined which features of an author's profile are most effective in mimicking the behaviour of expert users. We would now like to look at the effectiveness of considering each message (or in the SIGIR case, each paper) independently, looking just at the characteristic of the message. Again, before looking at the contribution of any single feature which we have identified as being of possible benefit to our re-ranking strategy, we must first look at other state-of-the-art approaches. We will not take into account any features of the author of each paper, instead looking just at the structure of the citation graph itself.

Gómez et al. (2008) have adapted the h-index of Section 4.1.2.1 to the web forum scenario. This scenario is very similar to the one we have been researching, and so it seems highly appropriate to look at the effectiveness of this implementation in our context. Gómez et al. study the threaded conversation which takes place within the Slashdot[6] forums.

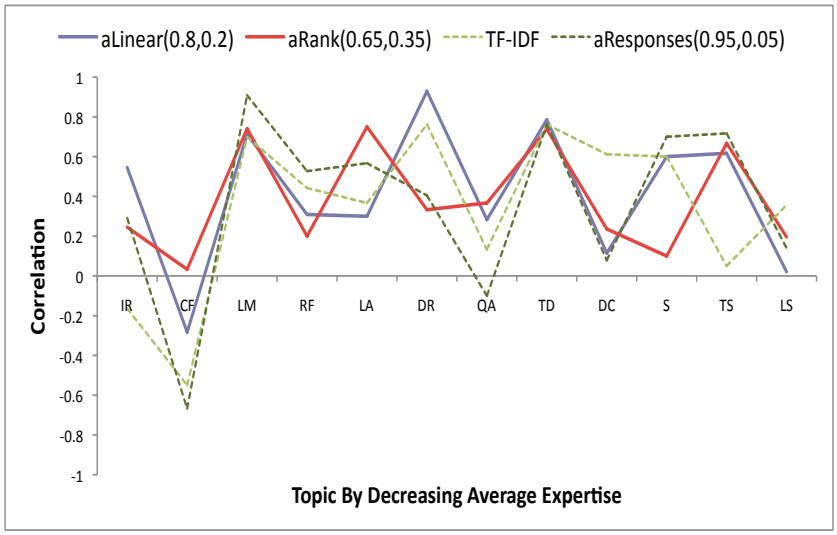
Like all forums, the form which these threads take is almost identical to our own SIGIR corpus. Users post an article with a short description. Other users within the community are then able to read and comment on this posting, with the replies taking a threaded structure. By viewing the threads as a tree, with messages radiating out from the original post, Gómez et al. are able to visualise the Slashdot corpus as a forest of **radial trees**. The original post forms the central node or root of the tree. Direct replies to this post appear at the first nested level; replies to these replies appear at the second nested level etc. This structure can be seen in Figure 80.

In order to measure the controversy or impact of a post, Gómez et al. propose a modified version of the h-index (which we shall call *h-Slash*) defined as follows:

[6] http://www.slashdot.com



(a) Performance on SIGIR_Txt



(b) Performance on SIGIR_Comb

Figure 79: Comparisons of the author-feature based re-rankings of the TF-IDF baseline from the two Lemur indexes, with the experts' rankings. We show here the performance of each of the two combinations of author features 'aLinear' and AuthorRank 'aRank', as well as the top performing single feature 'average responses'. Each measure also shows its respective optimal (α,β) weights. Lastly we show the TF-IDF baseline.

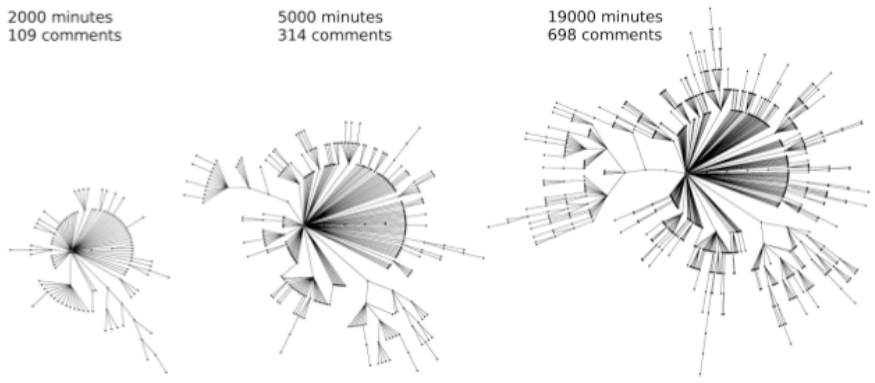


Figure 80: An example of radial tree structure corresponding to a controversial post related to Windows and Linux which received a total of 982 comments. The title of the post is "Can Ordinary PC Users Ditch Windows for Linux?". Figures show three snapshots at different times (Gómez et al., 2008).

Definition 7 *Given a radial tree corresponding to a discussion thread and its comments organised in nesting levels, the* ***h-Slash*** *(h) of a post is the maximum nesting level i which has at least $h > i$ comments, or in other words, $h + 1$ is the rst nesting level i which has more than i comments.*

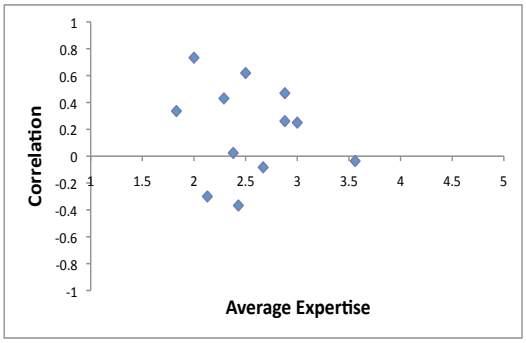
They note that a great many posts will have the same h-slash, and so a method of prioritising these messages is required. In order to rank posts with tied h-slashes, Gómez et al. give priority to those messages which reach a certain h-slash with less comments. Thus, for a post i the following ranking formula is used:

$$r_i = H_i + \frac{1}{C_i} \tag{33}$$

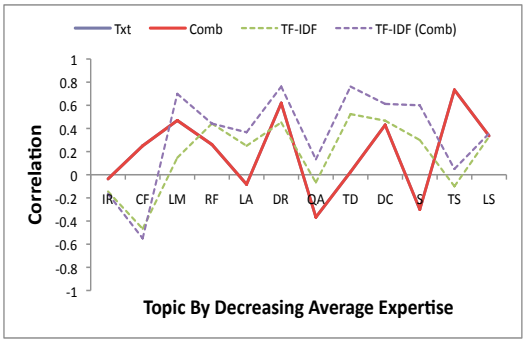
where H_i is the h-slash for post i, and C_i is the number of comments created on i in order to reach H_i.

We use Equation (33) to rank the papers within the SIGIR corpus. By considering each paper as the root of its own tree, we are able to then recreate the situation proposed above for the Slashdot forums. The h-slash value of a paper is the maximum nested level i at which citations of this paper have less than i citations. Again, we take into account the number of citations in total (C_i).

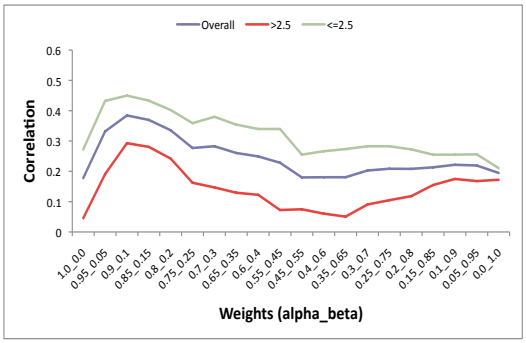
Ranking the messages by h-slash alone produces a positive increase which is significant in both indexes. The h-slash measure is a more link-based measure than the author specific measures, most similar to the m-index. Not alone does it take into account the effect of citations within the h-core of a paper, but gives additional credit to papers which are more seminal. A paper which is highly cited by highly cited papers will have a higher h-slash than a paper cited by many more low citation papers. The increase in



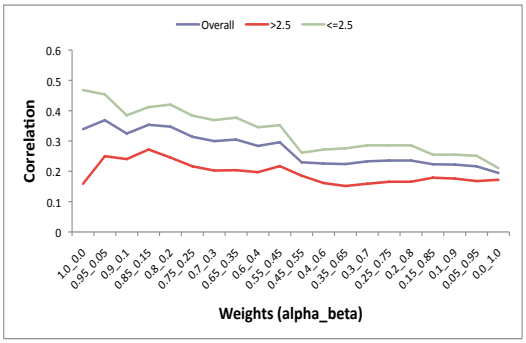
(a) 'h-Slash' correlation with experts' ranking



(b) Correlation of 'h-Slash' re-ranking with experts' ranking



(c) Experts' ranking correlation within the SIGIR_Txt index



(d) Experts' ranking correlation within the SIGIR_Comb index

Figure 81: Comparisons of the h-Slash re-rankings of the TF-IDF baseline from the two Lemur indexes, with the experts' rankings.

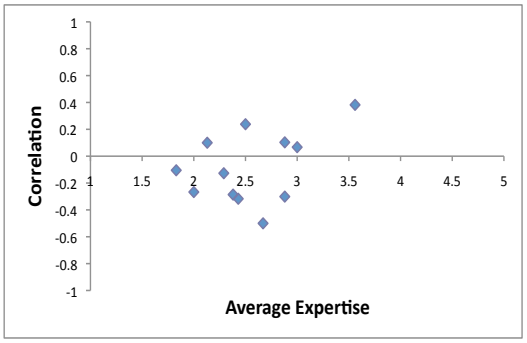
correlation from 0.04 to 0.29 within the SIGIR_Txt index between the h-slash re-ranking and the experts' ranking is 0.25 ($p = 0.078$), achieved by setting $\alpha = 0.9$ and $\beta = 0.1$. An increase from 0.16 to 0.27 is also achieved within the SIGIR_Comb index by setting $\alpha = 0.85$ and $\beta = 0.15$, an increase which again is significant ($p = 0.073$). In the case of the SIGIR_Txt index, this increase in correlation causes an insignificant raise in AP from 0.54 to 0.60 ($p = 0.289$). This is not the case in the SIGIR_Comb index however, where a significant fall in AP is seen from 0.61 to 0.50 ($p = 0.062$).

While the h-slash of messages is effective in raising the correlation of all topics in the SIGIR_Txt index, the lower expertise topics within the SIGIR_Comb index see a slight deterioration in correlation. It would seem that the h-slash measure is better at emulating the rankings of higher expertise raters, than those of lower expertise. Again, like the m-index, the h-slash measure gives credit to a paper if it is cited by many highly cited papers. In doing so, it also intrinsically favours older papers within the corpus which have had more time to accrue citations.

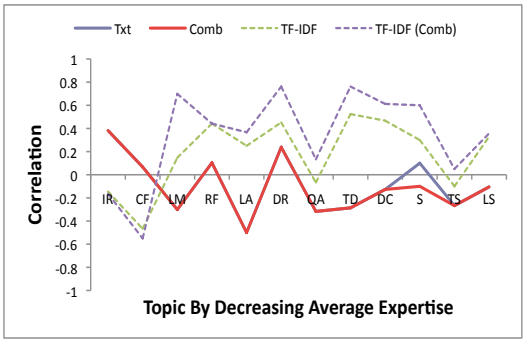
6.6 Calculation of Message Feature Contributions

Many of the features we will look at are closely related to the work which has gone before in the field of forum and news-group search. Xi et al. (2004) identified a number of features of postings within the news-group setting, a subset of which we have adapted for our own work. Justification for the use of these features is shown in both the work of Xi et al., as well as earlier work by Fiore et al. (2002) on the behaviour of authors within the news-group context. As we have seen, this context may be extended to the citation and annotation contexts quite easily. We shall examine the log-values of each of the features we are interested in; this is necessary in order to prevent near-exclusion of features due to normalisation. The features which we shall be looking at are as follows:

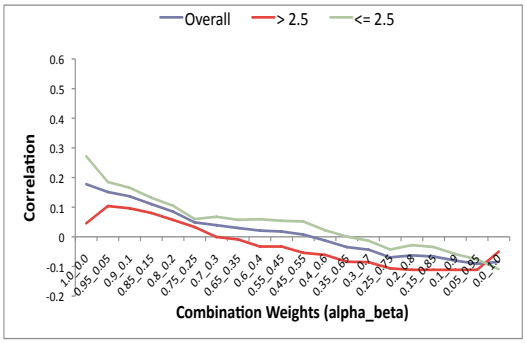
- **Message/Citation Words:** This is the number of words which are created by an author in citation of a previous work. Due to difficulties with the download and extraction of PDF documents, it is preferable to use the number of non-whitespace characters in place of a word count. We also use the term *paper* interchangeably with message, since each message in the context of SIGIR is in fact a paper which contains the citation-context we are interested in.
- **Average Thread Words:** This is the average number of words per message/citation within the containing thread of the message of interest. This is of interest as it provides a vague idea as to the amount of information being added on average per author.
- **Thread Words:** This is the total number of words contained in the thread which this message is found in. Again, we take the count of non-whitespace characters for reasons explained above.
- **Message Depth:** The depth at which the message/citation of interest is found within its containing thread. This depth is indexed from the earliest post, and begins at zero (i.e. the root message/paper is found at depth zero within a thread of length one.). We take the inverse of the log of message depth, since the lower in a thread the message is found, the less information is can claim any credit/involvement with.
- **Thread Length:** This is the maximum depth which a thread grows to. This maximum may be greater than the length of the branch of a thread in which a message is found (i.e. a message may cite a paper and receive two citations which themselves receive no citations. Another citing message/paper however, may receive one citation that then receives citations. In this case the thread length is three.).



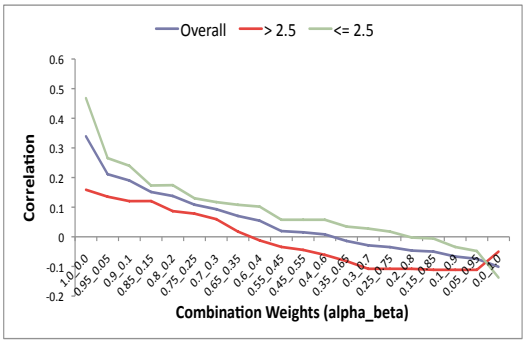
(a) 'Message words' correlation with experts' ranking



(b) Correlation of 'message words' re-ranking with experts' ranking



(c) Experts' ranking correlation within the SIGIR_Txt index



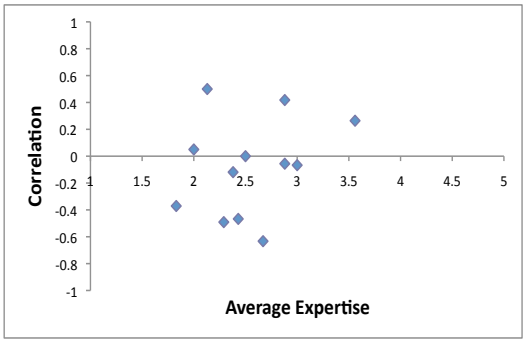
(d) Experts' ranking correlation within the SIGIR_Comb index

Figure 82: Comparisons of the 'message words' re-rankings of the TF-IDF baseline from the two Lemur indexes, with the experts' rankings.

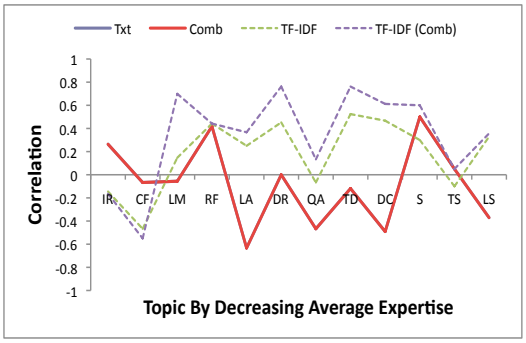
Looking first at the contributions of the message word count to the re-ranking of the TF-IDF baseline, we see from Figure 82(a) that there is no real correlation between average expertise and the re-rankings achieved by the 'message words' feature. While the feature does provide a small boost in correlation within the higher expertise topics of the SIGIR_Txt index for $\alpha = 0.95$ and $\beta = 0.05$, this increase of 0.06 from 0.04 to 0.1 is not significant ($p = 0.161$). The inclusion of feature information in conjunction with TF-IDF on the SIGIR_Comb index is universally detrimental. Again, the effect may be dampened by the heuristic choice of citation-context limits. The corresponding AP values for the SIGIR_Txt index sees an increase from 0.54 to 0.58, however this is not significant ($p = 0.156$).

If the heuristic choice of citation-context length was the reason for the insignificant increase ascribed to the 'message words' feature, we would expect to see a similar situation in the combinations obtained from re-ranking with the 'average thread words' feature. On the contrary, this feature appears to perform far better than single message word counts[7]. A slight positive correlation may be observed in Figure 83(a), and we

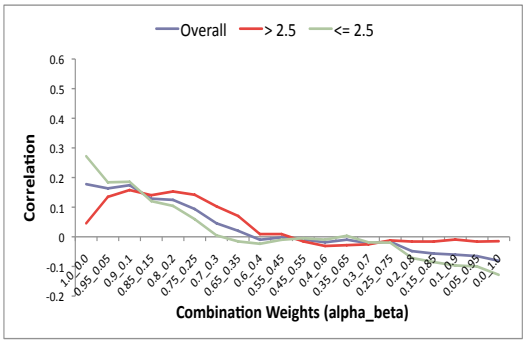
[7] An important difference between the 'average thread words', and 'message words' features is in the construction of the message statistics. A message will only have words (given by citation-context) if it cites a previous paper in the corpus. This may be why the older topics, such as *'language modelling'* (LM) and *'link analysis'* (LA) are negatively effected in a strong way by inclusion of the 'message words'



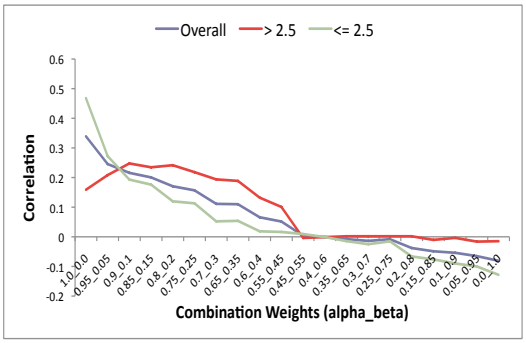
(a) 'Average thread words' correlation with experts' ranking



(b) Correlation of 'average thread words' re-ranking with experts' ranking



(c) Experts' ranking correlation within the SIGIR_Txt index



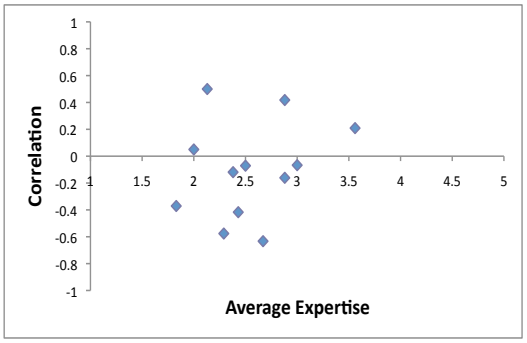
(d) Experts' ranking correlation within the SIGIR_Comb index

Figure 83: Comparisons of the 'average thread words' re-rankings of the TF-IDF baseline from the two Lemur indexes, with the experts' rankings.

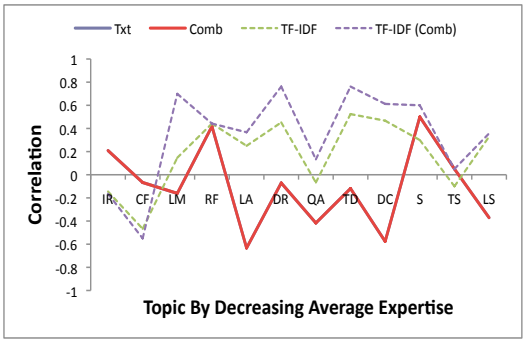
see that with the exception of the 'spam' (S) topic, all the lower expertise topics show slight to high negative correlation between the re-rankings produced for these topics by the 'average thread words' feature and our experts' rankings. Indeed, while the inclusion of the feature information is universally detrimental on both indexes for the lower expertise topics, we can see that it has a large positive effect on the higher expertise topics. The correlation is increased significantly ($p = 0.075$) from 0.04 to 0.15 within the SIGIR_Txt index, and 0.16 to 0.25 within the SIGIR_Comb index. The increase within the SIGIR_Comb index is not however significant ($p = 0.262$). Both increases occur when $\alpha = 0.9$ and $\beta = 0.1$. These same values of α and β see AP within the SIGIR_Txt index raise slightly from 0.54 to 0.56, and within the SIGIR_Comb index there is a fall from 0.61 to 0.59. Neither of these changes are significant however with $p = 0.283$ and $p = 0.160$ respectively.

Looking at the thread word count of a message's containing thread, we see from Figure 84(a) that there is again a slight positive correlation between the re-ranking of TF-IDF by the 'thread words' feature, and our experts' rankings. The inclusion of the feature information sees a significant increase in correlation within the SIGIR_Txt index ($p = 0.066$), increasing from 0.05 to 0.15. There is a increase within the SIGIR_Comb

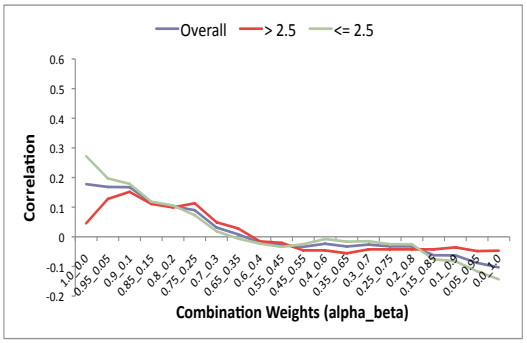
feature information.



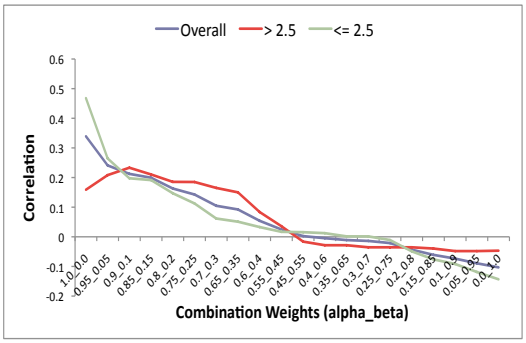
(a) 'Thread words' correlation with experts' ranking



(b) Correlation of 'thread words' re-ranking with experts' ranking



(c) Experts' ranking correlation within the SIGIR_Txt index



(d) Experts' ranking correlation within the SIGIR_Comb index

Figure 84: Comparisons of the 'thread words' re-rankings of the TF-IDF baseline from the two Lemur indexes, with the experts' rankings.

index also from 0.16 to 0.23, however this increase is not significant ($p = 0.298$). While AP increases by 0.54 to 0.57 within the SIGIR_Txt index, and decreases from 0.61 to 0.58 within the SIGIR_Comb index, not surprisingly, neither of these changes are significant ($p = 0.406$, and $p = 0.117$ respectively). In both cases, $\alpha = 0.9$ and $\beta = 0.1$.

The inclusion of thread information, either directly through the 'thread words' feature, or slightly more indirectly through the 'average thread words' feature, appears to be significantly positive within the SIGIR_Txt index if not the SIGIR_Comb index. We conclude that the inclusion of thread information in this context provides a definite boost to correlation, though not as much as the inclusion of citation-context text. Its effect is somewhat nullified by the inclusion of this information however, as the significance is lost within the SIGIR_Comb index.

Looking now at the position of the message within the thread, we can see that the 'message depth' feature not alone provides a large boost to correlation within the SIGIR_Txt index (figure 85(c)), but in fact increases the correlation within the higher expertise topics to a level higher than that of the overall correlation. Surprisingly, this increase in correlation, setting $\alpha = 0.85$ and $\beta = 0.15$, from 0.05 to 0.22 is not significant ($p = 0.119$). This does however cause a fall in AP from 0.54 to 0.47 which fortunately

again is not significant ($p = 0.153$). Turning to the SIGIR_Comb index we see that again the inclusion of feature data proves beneficial, increasing correlation from 0.16 to 0.28. Neither increase, nor the corresponding fall in AP from 0.61 to 0.48 is significant ($p = 0.194$; $p = 0.139$). Despite this, the depth of a message within a thread does seem to provide a powerful indication of its value or importance.

The last feature of interest is the 'thread length' feature, comprising of information about the size of the thread in which a message/citation is found. Figure 85(f) shows that despite the positive correlation between expertise and re-ranked lists depicted in Figure 85(e), nearly all per-topic correlations though increasing with expertise, are in fact negative. The feature proves wholly detrimental to both indexes, bringing the correlation down in all cases. The length of a thread does not seem to provide any useful information on the impact of the messages within the thread. It should be noted however, that the effect of thread length may be curtailed due to the fixed window size used in our experiments. No paper can be contained in a thread of any great length due to this window. Also, due to the nature of out corpus, papers can only be published at a fixed time-point.

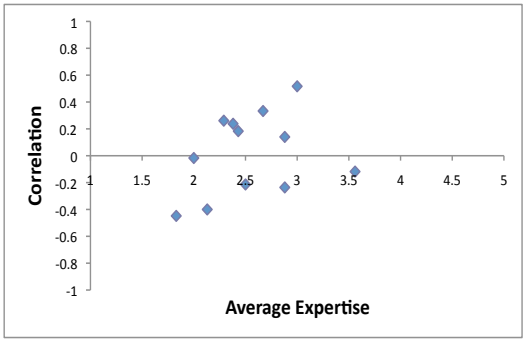
Of the features we have examined, the most effective in improving correlation between experts' ranking and the feature's re-ranking of TF-IDF seem to be the features which incorporate contextual information about the message, rather than relying on just the message itself. The boost provided by any single feature however is not as significant as that provided by the h-slash re-ranking. The h-slash measure, while ignoring the 'message depth' aspect (since it effectively assumes that every message is at depth zero), incorporates more information about the structure of the thread a message is found in than just the thread length. We now look at combining our features in such a way as to take similar advantage of the message context.

Table 24: Optimal combinations achieved for single message feature measures.

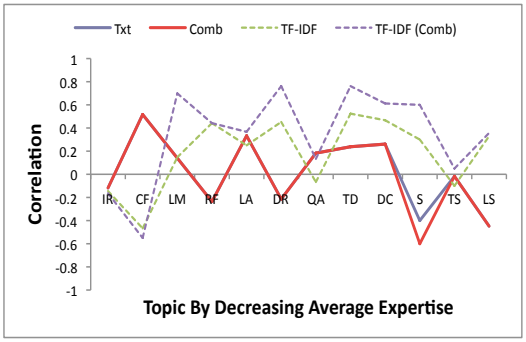
Feature	*SIGIR_Txt*				*SIGIR_Comb*			
	α	β	Corr.	A.P.	α	β	Corr.	A.P.
Message Words	0.95	0.05	0.06(0.02)	0.58(0.04)	1.00	0.00	0.16(-)	0.70(-)
Avg. Thread Words	0.90	0.10	**0.15(0.11)**	0.56(0.02)	0.90	0.10	0.25(0.09)	0.59(-0.02)
Thread Words	0.90	0.10	**0.15(0.10)**	0.57(0.03)	0.90	0.10	0.23(0.07)	0.58(-0.03)
Message Depth	0.85	0.15	0.22(0.17)	0.47(-0.07)	0.85	0.15	0.28(0.08)	0.48(-0.13)
Thread Length	1.00	0.00	0.05(-)	0.65(-)	1.00	0.00	0.16(-)	0.70(-)

6.6.1 Combination of Message Features

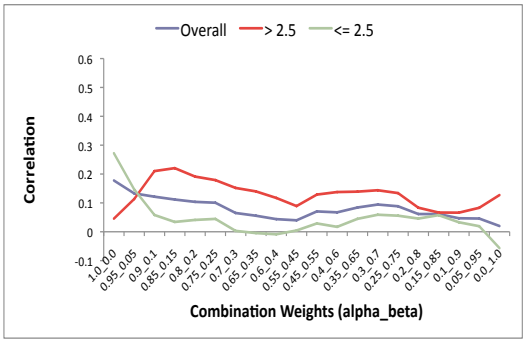
We have looked at the impact of each of our message features alone. We would now like to combine these different features gaining the benefit of each. The first three of



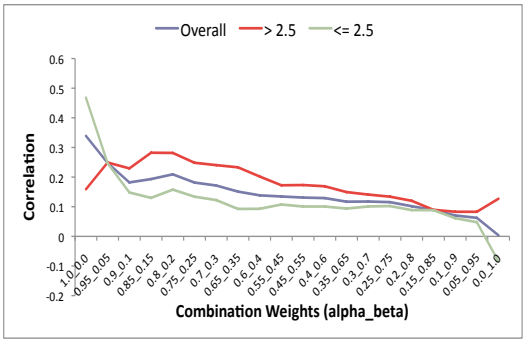
(a) 'Message depth' correlation with experts' ranking



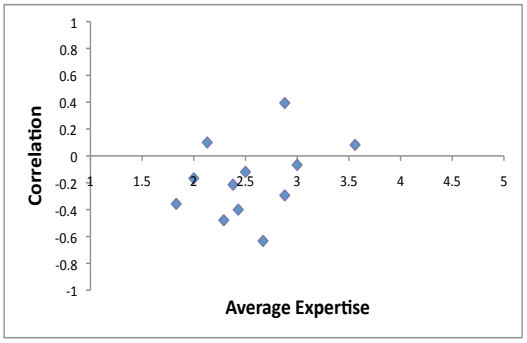
(b) Correlation of 'message depth' re-ranking with experts' ranking



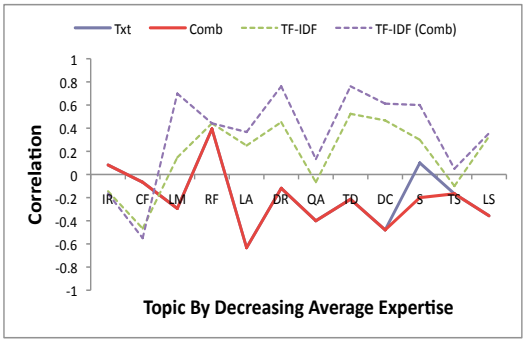
(c) 'Message depth' ranking correlation within the SIGIR_Txt index



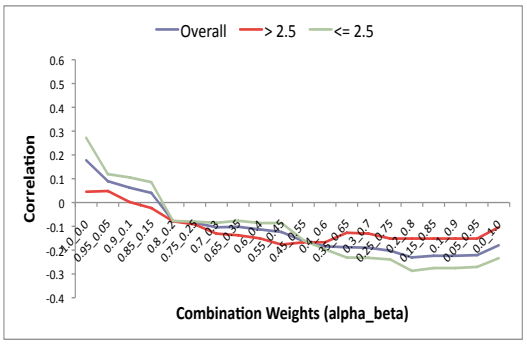
(d) 'Message depth' ranking correlation within the SIGIR_Comb index



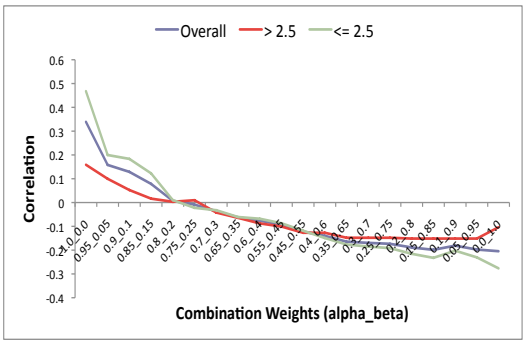
(e) 'Thread length' correlation with experts' ranking



(f) Correlation of 'thread length' re-ranking with experts' ranking



(g) 'Thread length' ranking correlation within the SIGIR_Txt index



(h) 'Thread length' ranking correlation within the SIGIR_Comb index

Figure 85: Comparisons of the 'message depth' and 'thread length' re-rankings of the TF-IDF baseline from the two Lemur indexes, with the experts' rankings.

these features look at the amount written within a single message/citation as well as the containing thread, taking advantage of the contributions of the message in terms of the conversation around it. There is however no sense of when this message appears within the thread, no real temporal context nor credit for the amount of discussion which comes after this message. The last two features play the opposite role in identifying the context of the message within its surrounding conversation, but not the size of the message with respect to the thread.

Table 25: Optimal feature weights for linear combinations of the message features. Topics are listed in order of decreasing expertise, and are divided by a dotted line representing the two classes of higher and lower expertise.

(a) SIGIR_Txt weights

Topic	*Weights*					Corr.
	log(MW)	log(TW)	log(MD)	log(TL)	log(ATW)	
IR	1.00					0.382
CF	0.10	0.05	0.45		0.40	0.567
LM		0.20	0.40		0.40	0.573
RF	0.05	0.10	0.70	0.10	0.05	0.673
LA	0.15		0.85			0.550
DR	0.65	0.05	0.15	0.15		0.357
QA	0.10		0.90			0.217
TD	0.15		0.20		0.65	0.517
DC	0.10	0.35	0.45	0.10		0.511
S			0.60		0.40	0.600
TS	0.05	0.15	0.65		0.15	0.714
LS	0.95		0.05			-0.105

(b) SIGIR_Comb weights

Topic	*Weights*					Corr.
	log(MW)	log(TW)	log(MD)	log(TL)	log(ATW)	
IR	1.00					0.382
CF	0.10	0.05	0.45		0.40	0.567
LM		0.20	0.40		0.40	0.573
RF	0.05	0.10	0.70	0.10	0.05	0.673
LA	0.15		0.85			0.550
DR	0.65	0.05	0.15	0.15		0.357
QA	0.10		0.90			0.217
TD			0.35		0.65	0.477
DC	0.10		0.90			0.309
S		0.10	0.65		0.25	0.800
TS		0.20	0.40	0.10	0.30	0.619
LS	0.95		0.05			-0.105

Table 26: Optimal weights for linear combination of author features across topics

Topics	*Weights*				
	log(MW)	log(TW)	log(MD)	log(TL)	log(ATW)
Top	0.05	0.45	0.50		
All			0.40	0.05	0.55

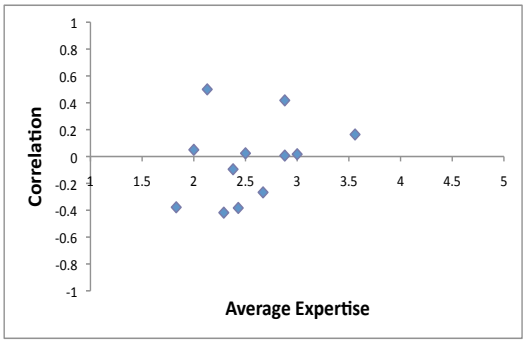
In combining these features, we first look to use the simplest method of linear combination. The five features of a message are combined in a weighted manner, with the optimal weights for each topic given in Table 25. In both Table 25(a) and Table 25(b) we can see that the 'message depth' feature plays a very important role in nearly all topic feature combinations. From the discussion above (and most specifically Figures 85(c) and 85(d)) we have seen that this feature did provide a substantial increase in correlation between the higher average-expertise topics' rankings, and its own re-ranking of the TF-IDF baseline. In fact in all of the higher expertise topics, the 'message depth' feature receives the greatest weighting.

In both indexes, the weights used to optimise the correlation of each highly ranked topic to the experts' ranking remain the same. While 'message words' plays a small part, at the two extremes of expertise (*'image retrieval'* (IR) and *'latent semantic [indexing/analysis]'* (LS)) it is the greatest/only weighted feature. 'Message words' on a per paper basis (as opposed to the situation in MessageRank where every citation is treated as a message in its own right) contains a combined word-count of all the citation-contexts created by a paper in referencing other papers. As such, if a paper sites a large number of other papers, especially other SIGIR papers, it will have a large 'message words' feature. It may be the case that with the IR and LS topics, the papers which are ranked highly by our experts are papers which happen to have cited a large number of SIGIR papers.

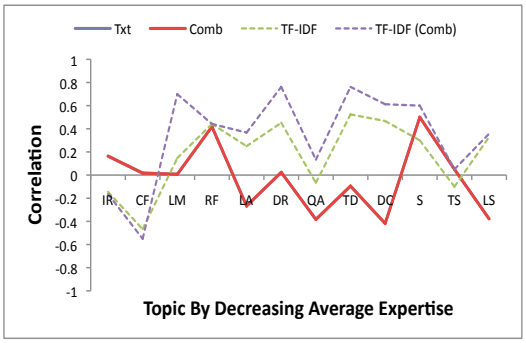
The 'thread length' feature is of little benefit in just three topics with relevance feedback (RF) being the only higher expertise topic to provide any weight to it at all. In this context it would appear that being cited by many papers, rather than by a few papers which are cited many times etc. is of more benefit. The 'average thread words' feature is of greater importance to the lower expertise topics on average, but only slightly.

Turning now to the weighting combinations which give the best correlations across topics, we see that the influence of the 'message depth' feature is indeed prevalent in all topic expertise levels. Table 26 shows the weights which should be used to create re-rankings of the TF-IDF baseline which most highly correlate with our experts' ranking. We can see that for both the higher expertise (Top) topics, and overall (All), the feature

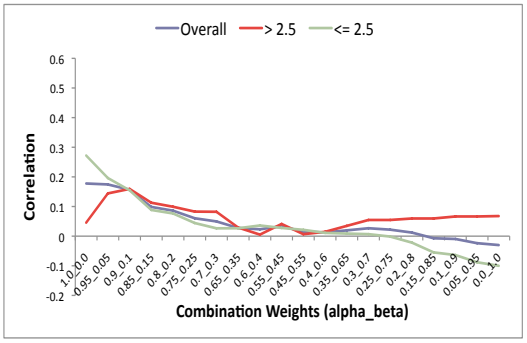
plays an important role. In the case of the higher expertise topics, it is the most heavily weighted feature. This may reflect the way in which more expert raters will consider not just the paper itself when ranking, but also the past work in which the research is grounded.



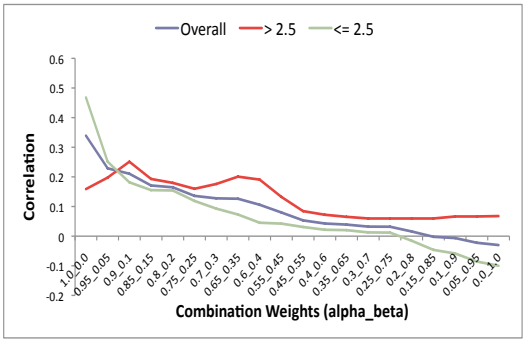
(a) Linear message feature combination's correlation with experts' ranking



(b) Correlation of the linear message feature combination's re-ranking with experts' ranking



(c) Experts' ranking correlation within the SIGIR_Txt index



(d) Experts' ranking correlation within the SIGIR_Comb index

Figure 86: Comparisons of the linear message feature re-rankings of the TF-IDF baseline from the two Lemur indexes, with the experts' rankings.

'Thread length' as we have said will not vary as much as the depth at which a message is found, and indeed plays no part in the weighting combination for higher expertise topics, and only a small part in the optimal weighting combination for all topics. For the higher expertise topics, the characteristics of a message itself are more useful than average thread information. The combination of 'message words' and 'thread words' features take the place of the 'average thread words' feature, which is highly weighted for lower expertise topics.

Taking this linear combination of features as a re-ranking method, and applying it to the TF-IDF baseline as we have done with each of the features singly yields an increase in correlation with the experts' ranking as good as any single feature. That is with the exception of 'message depth' which provided a bigger boost to correlation, but not significantly. Setting $\alpha = 0.9$ and $\beta = 0.1$ on both indexes show increased correlation within the higher expertise topics which we are interested in. The increase in correlation within the SIGIR_Txt index from 0.05 to 0.16 is significant ($p = 0.058$), along with an

increase in AP from 0.54 to 0.55 which is not ($p = 0.383$). The SIGIR_Comb index also experiences an increase in correlation from 0.16 to 0.25, but as in most cases with the single features, this increase is not significant ($p = 0.246$). There is however a significant fall in AP ($p = 0.058$) which is reduced from 0.61 to 0.57.

From these results we see that the ability of query-independent message features to increase correlation with expert users' rankings is significant. While it is also helpful in the case of the SIGIR_Comb index, where citation-context has been added into the documents, the increase is no longer significant. This loss of significance may be due to the fact that less-noisy information about the paper (i.e. the citation-contexts of citing papers) has already been taken into consideration in the original TF-IDF ranking. In the next section we shall attempt to look beyond just the information about a paper, and take into account both its author, and more specific information on who has been citing the paper.

6.7 Calculation of MessageRank Weights

Finding the optimal weights for a linear combination of the message features again allows us to see how well these features can perform in re-ranking the TF-IDF baseline to improve correlation with our experts' ranking. With MessageRank we not alone take into account whether a paper/message has been cited, or how large its own citation-context is, but instead look to incorporate additional information on who has been citing it. Recall the MessageRank formula is of the form:

$$M_R = A_R * \{\frac{2 \log M_w}{\log T_w * \log T_a} * [\log T_l - \log M_d]\} + \tau * [\sum_{x=1}^{n} \frac{A_{R_x}}{e^{d_x}}] + (1-\tau) * [\sum_{y=1}^{m} \frac{A_{R_y}}{e^{d_y}}] \quad (34)$$

Each of the single features is again incorporated to aid in the re-ranking of the TF-IDF baseline; the AuthorRank (A_R) of this paper's authors; message words (M_w); thread words (T_w); average per-message words in a thread (T_a); thread length (T_l); and finally message depth (M_d).

In addition to this, MessageRank takes into account who else is involved in the conversation/thread with this message's author. To do this we look at the containing thread and include the AuthorRank of the authors who have replied directly to this message (cited this paper), and the AuthorRank, A_{R_x} of the authors who have replied to that reply. We limit ourselves to a nesting depth of two, since this mimics the friend-of-a-friend analogy of Watts and Strogatz (1998). We discount the value by dividing by e^{d_x}, where d_x is the nesting depth of this author. We would also like to include information on the authors occurring above this message in the thread, since this will

help us in some way to gauge the quality of the conversation. In much the same way as we have differentiated between 'replied' and 'started' threads in AuthorRank, we differentiate between those authors occurring above and below this message within the thread. We use τ to do this. After performing a grid-search, we see the optimal value for $\tau = 0.9$ in the case of SIGIR_Txt, and $\tau = 0.85$ for SIGIR_Txt as shown in Table 27.

Table 27: Optimal weights for τ across topics. This is the weight given to author appearing above the message of interest. Replies/citations to this message recieve a weight of $(1 - \tau)$.

(a) Higher Expertise Topics

Index	*Topic*					
	IR	CF	LM	RF	LA	Top
SIGIR_Comb	0.65	0.00	0.85	0.55	0.90	0.85
SIGIR_Txt	0.60	0.00	1.00	0.45	0.90	0.90

(b) Lower Expertise Topics

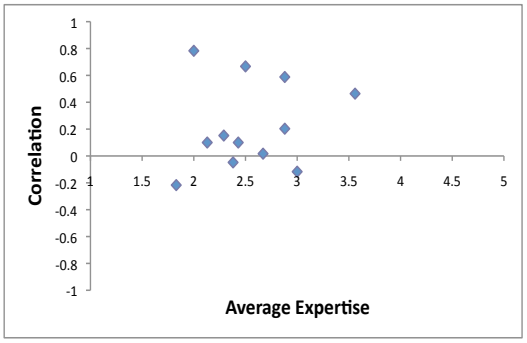
Index	*Topic*							
	DR	QA	TD	DC	S	TS	LS	All
SIGIR_Comb	0.55	0.00	1.00	0.05	0.35	0.30	0.00	0.80
SIGIR_Txt	0.55	0.00	1.00	0.05	0.35	0.45	0.00	0.90

Looking first at the higher expertise topics, we notice that in all but the collaborative filtering (CF) topic, τ is greater than 0.5. This means that the influence of the messages/citations above the message of interest are of greater importance than those below; in other words, the conversation present before this message plays a more significant role in creating an optimal re-ranking than any of the replies or later messages. While it is not clear why earlier message/papers within the *'link analysis'* (LA) topic are given so much weight, the situation in the *'language modelling'* (LA) and *'topic distillation'* (TD) tasks may be explained by the fact that both topics containing very heavily cited papers within the set of expert-ranked documents[8].

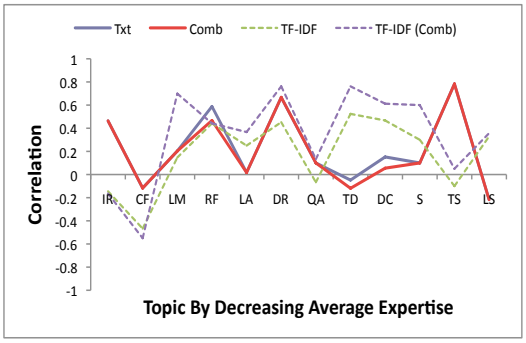
It is not clear why there is a divide in how the higher and lower expertise classes of topics apportion the influence of earlier and later papers. The average optimal value for τ in the higher expertise topics is 0.59, while in the lower expertise topics it is 0.32. It would appear that perhaps experts of higher expertise will take into account factors such as a paper's grounding in past research when ranking by importance. This would intuitively appear to make sense, as a person with little expertise within a topic field would not know much about past research and could therefore not factor it in when

[8]In the case of 'language modelling', this is *A Language Modeling Approach to Information Retrieval.* In 'topic distillation' it is the paper *Improved Algorithms for Topic Distillation in a Hyperlinked Environment.*

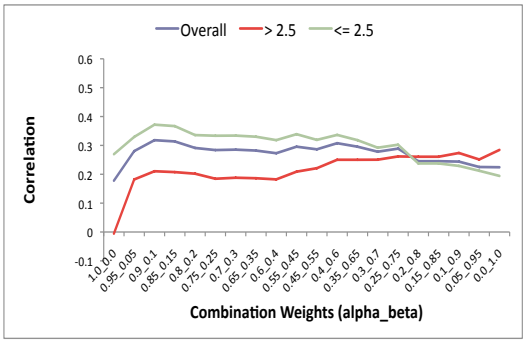
deciding on an appropriate ranking.



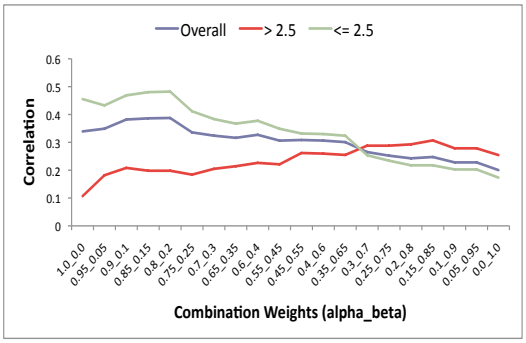
(a) MessageRank's correlation with experts' ranking



(b) Correlation of MessageRank's re-ranking with experts' ranking



(c) Experts' ranking correlation within the SIGIR_Txt index



(d) Experts' ranking correlation within the SIGIR_Comb index

Figure 87: Comparisons of MessageRank's re-rankings of the TF-IDF baseline from the two Lemur indexes, with the experts' rankings.

Using MessageRank alone to re-rank the TF-IDF baseline results in a boost in correlation which in the case of the SIGIR_Txt index is significant ($p = 0.0814$). Setting $\alpha = 0.0$ and $\beta = 1.0$ increases correlation from -0.02 to 0.38. While this results in correlation is twice as good as that of the optimal linear combination of features (Figure 86(c)), there is no significant difference between the two ($p = 0.31$). There is however a significant reduction in AP which falls from 0.56 to 0.24 ($p = 0.033$). In the case of the SIGIR_Comb index, setting $\alpha = 0.15$ and $\beta = 0.85$ increases correlation four-fold from 0.10 to 0.40. Despite this, the resultant increase is not significant ($p = 0.1100$). Nor is it a significant improvement on the best linear combination of features, despite having a correlation of nearly double that of the linear combination ($p = 0.295$). The corresponding fall in AP however from 0.70 to 0.21 is ($p = 0.008$), meaning that in the case of both indexes, there is a significant fall in AP due to the re-ranking of documents.

This fall in AP is due to the spreading out of relevant documents within the returned result set. This should not be considered a major concern, since the documents which were chosen for the experts to rank were not always from the top documents[9].

[9]By this we mean that documents which appeared in both the SIGIR citation PageRank graph, and the Google Scholar list of returned document may not have occurred highly in both. See page 129 for

It should be noted that it does not really make sense to use MessageRank alone to re-rank the papers returned by the TF-IDF baseline. MessageRank is measured on the citations of a paper, not the paper itself and is used to take advantage of additional contextual information about the documents/papers being ranked. In order to perform the comparisons above, we have propagated the MessageRank scores of the citing documents (those which are directly citing the document to be ranked) onto the document of interest. We then take the average of these values as the MessageRank of the document. In the next section, we utilise MessageRank as it was originally intended - in conjunction with the initial ranking function, TF-IDF, and the AuthorRank of a document's author.

6.7.1 The Performance of MessageRank

We would like to compare the performance of MessageRank to the current state-of-the-art, as well as the straight linear combination which we have created in Section 6.6.1. To do so, we must create the full weighting scheme, since we would not use MessageRank alone to re-rank a returned results set. As stated, MessageRank incorporates information about the comments which are made on a document, allowing us to take into account not just the original document, but the network of comments and meta-data about the document.

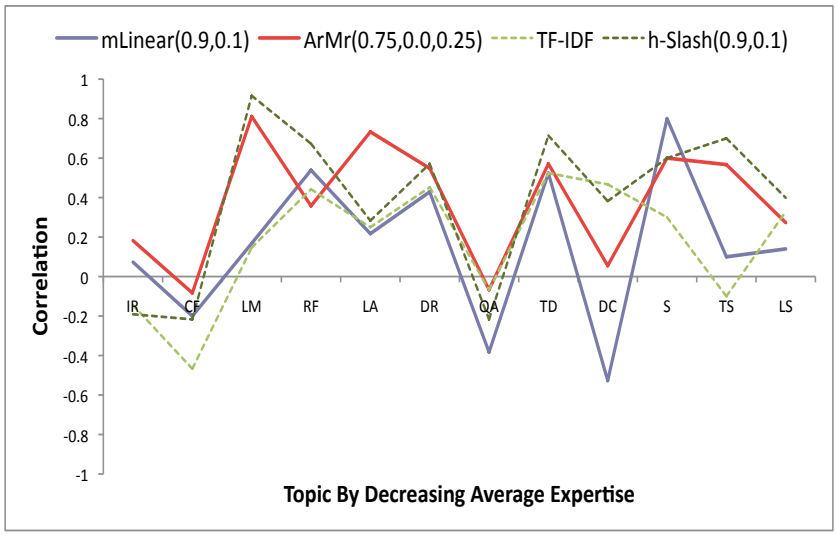
Before adding in this information, we must first retrieve the documents of interest. As we have been doing up to this point, we continue by using the TF-IDF method. Once we have our initial ranking of documents, we now re-rank this list of documents based on a combination of their TF-IDF score, the AuthorRank (A_R) of their authors, and the MessageRank, (M_R), of any comments/citations which have been made on the document. This is shown in Equation (35).

$$A_r M_r = \alpha * \text{TF-IDF} + \beta * A_R + \gamma * \frac{1}{n} \sum_{i=1}^{n} M_{R_i} \tag{35}$$

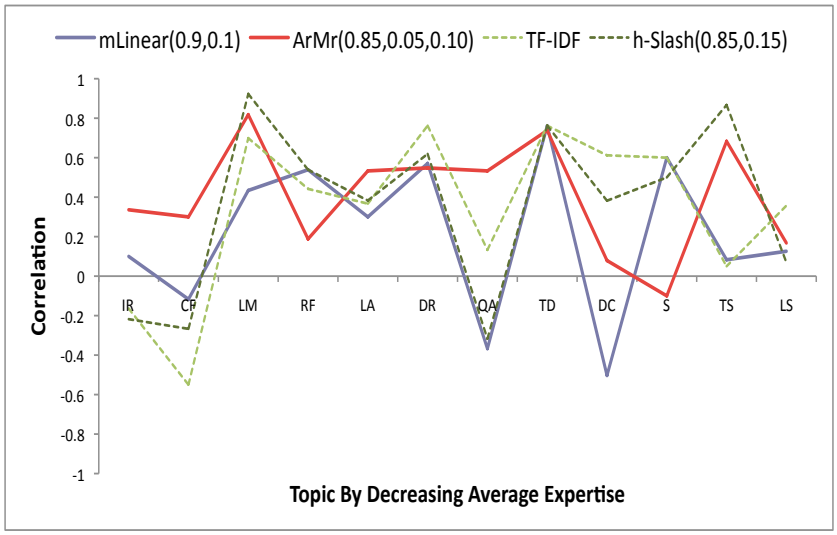
In order to see how well this ranking performs, we compare it to the rankings generated by the h-slash measure, and that of the linear combination 'mLinear'. These can be seen in Figure 88.

The performance of our MessageRank and AuthorRank algorithms boosts correlation with the experts' ranking significantly in comparison to the intial TF-IDF baseline. Using weights of $\alpha = 0.75$, $\beta = 0.0$ and $\gamma = 0.25$ on the SIGIR_Comb index produces and increase in correlation from 0.16 to 0.44. This correlation borders on significance ($p = 0.101$). The improvement fails however to be a significant improvement on the rankings provided by either the linear combination 'mLinear' ($p = 0.128$), or that of

more details on the selection of documents for expert ranking.



(a) Performance on SIGIR_Txt



(b) Performance on SIGIR_Comb

Figure 88: Comparisons of the message feature based re-rankings of the TF-IDF baseline from the two Lemur indexes, with the experts' rankings. We show here the performance of both the straight combination of message features 'mLinear', and the combination of TF-IDF, AuthorRank and MessageRank 'ArMr' as shown in Equation (35). Also included is the feature 'h-Slash'. Each measure also shows its respective optimal (α,β[,γ]) weights. Lastly we show the TF-IDF baseline.

the 'h-Slash' measure ($p = 0.209$). The h-Slash measure (being a variant of the h-index) takes into account just those papers which have received a sufficient number of citations. It does not however take into account any author information. Setting $\beta = 0.0$ in Equation 35 results in the same ignoring of author features. The difference between the two algorithms then is that MessageRank differentiates between nesting levels of comments as with h-Slash, but h-Slash does not penalise comments which are found at more deeply-nested levels.

The performance of our algorithms on the SIGIR_Txt index is more encouraging. Setting $\alpha = 0.85$, $\beta = 0.0$ and $\gamma = 0.15$ yields an increase in correlation with the experts' ranking of 0.36, significantly increasing the baseline TF-IDF correlation from 0.05 to 0.40 ($p = 0.023$). On this index, the performance of the A_rM_r algorithm is also significantly better than that of the linear combination 'mLinear' ($p = 0.091$). It would appear that when the citation-contexts are not added in to the index as part of their referenced document, the additional information provided by the author features are of more pronounced benefit. The benefit does not however extend to surpassing the h-Slash measure; in this case, the improvement generated by the A_rM_r algorithm on the TF-IDF baseline is not significantly better ($p = 0.248$).

The combination of AuthorRank and MessageRank does manage to raise the correlation with the experts' ranking in 4 of the 5 top expertise topics. As we have said, this is significant in the case of the SIGIR_Txt index leading us to believe that it is possible to better emulate the considerations of a more expert user by taking the associated message features into account. In both indexes it would appear that taking information about a documents author into consideration is not as useful as taking into account the social network around the author. This may be seen in the fact that neither of the optimal combinations of AuthorRank, MessageRank and TF-IDF give any additional weight to the AuthorRank of a document's author over that given within the MessageRank algorithm itself.

6.8 Comparisons with the SportsAnno Corpus

In order to test the robustness of our weights, and subsequently our algorithms, we would like to ensure that the combinations which we have trained on our extended SIGIR corpus are not specific to this corpus. To do this, we use the SIGIR corpus as our training data, and our SportsAnno corpus as a source of test data. This also overcomes a second issue with the SIGIR corpus; we had been using this as a substitute for a real-world corpus of annotated data. Now that we have been able to train on this data-set however, we may go back to the real-world SportsAnno corpus which was created as part of the FIFA World Cup 2006 experiments.

Using the weights which we have trained on the SIGIR corpus, we aim to improve the ranking of relevant documents returned in answer to a query by users of the system. To do this we must first establish the same requirements that we had for our SIGIR system.

6.8.1 Collection of a Ground-Truth

In order to create a ground-truth against which to compare our algorithm's performance, we asked a group of people to provide ratings for the comments which have been created within the SportsAnno corpus. This group consisted of 16 individuals, some of whom where familiar with the original system. It was ensured that no user however was ever asked to grade the quality of their own annotations.

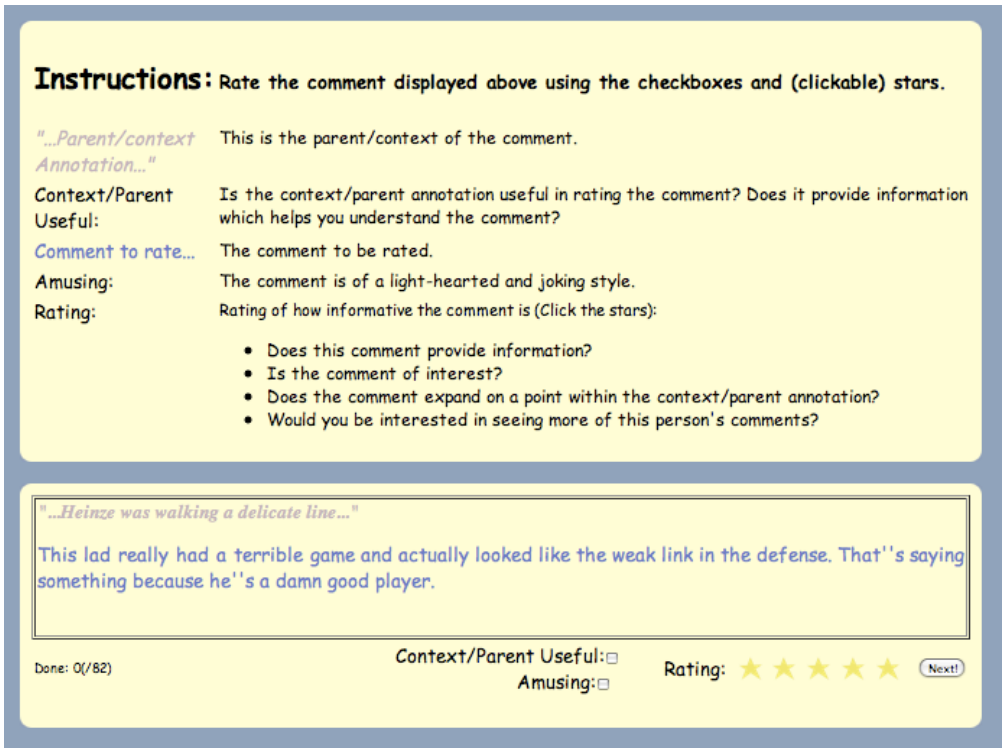


Figure 89: The interface presented to users when asked to provide a rating for each comment within the SportsAnno corpus.

Each user was required to evaluate the value or interest of a randomly selected subset of comments from the SportsAnno corpus. They did so by using the system shown in Figure 89. Each user was presented with a comment, along with some contextual information for this comment. This context was provided by the comment's parent. Recall that the SportsAnno system allowed users to comment directly, in-context, on written reports about matches within the FIFA World Cup 2006. If a comment was made directly on the report, and not in reply to a previous comment, then its parent became the selected text from within the report. This can be seen within Figure 89 where the parent is the phrase *"...Heinze was walking a delicate line..."*. This text is

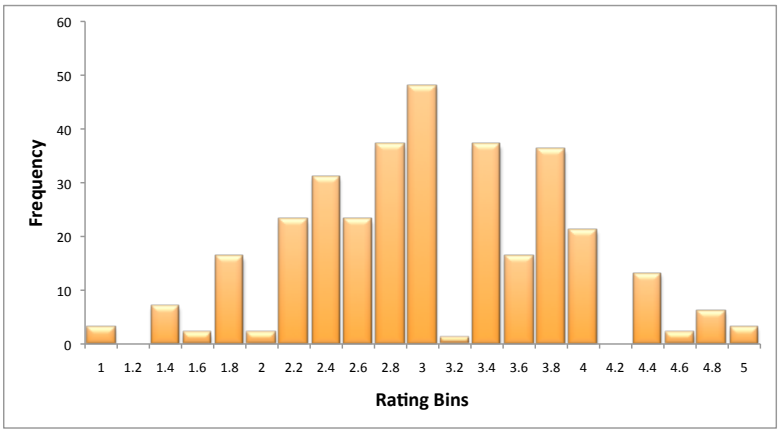


Figure 90: Histogram showing the ratings given to the 327 comments made in the SportsAnno corpus.

taken directly from a report. The comment which the user is required to rate is then shown directly below this.

Users were asked to rate comments on a scale of 1-5; 5 being a really useful comment; 1 means a comment is of no use. The usefulness/interest of comments was judged by users based on the following criteria:

- **Informativeness:** Does this comment provide information? Are facts stated that could be considered useful to another user? An example of this might be the comment *"I think he plays for Bayern Munich"*.
- **Interest:** Does this comment have something interesting to say? Is an opinion expressed that is of value to the community? By this we mean, is there evidence provided in justification of the expressed opinion?
- **Expansion:** Does this comment expand on the information or points made in it's parent quote/comment?
- **Personal Interest:** Would the user like to hear more of the opinions of this comment's author? This may be due the informativeness of a comment, or indeed just a personal choice.

Figure 90 shows the ratings that were given to the 327 comments which made up the SportsAnno corpus. Each comment was rated 5 times by 5 different users, and the average of these ratings was assigned to the comment. The comment distribution has a mean $\bar{x} = 2.974$ and variance $\mu = 0.664$.

Unlike the SIGIR corpus, all of the additional text provided by the comments is not already contained within the corpus. As such, the rating a comment receives reflects the quality of that comment directly. We also have exact information on the 'message

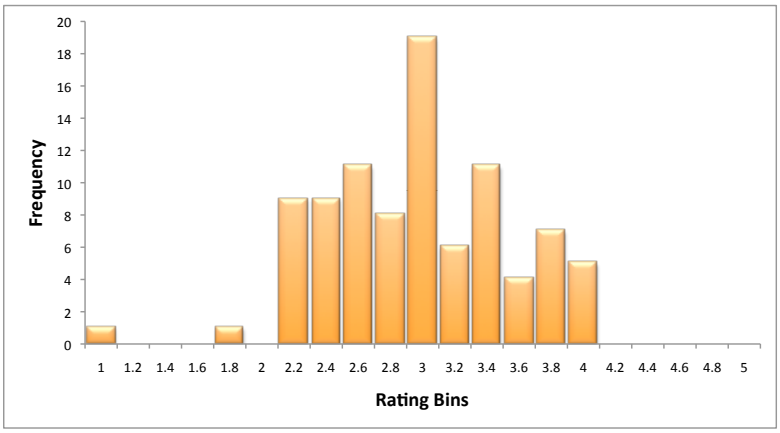


Figure 91: Histogram showing the ratings given to the 91 reports made in the SportsAnno corpus.

words' within each message. In order to create a ground-truth ranking of documents, we must make the assumption that the level of interaction around a document is a good indication of the quality. We also assume that this interaction provides a measure of its usefulness or interest to any users returned that document in answer to a query. In doing so, we are able to propagate the ratings received by direct in-context comments up to the document itself. By this we mean all comments which are made directly on the text, therefore being the head of any thread in which they are found. Of the 246 reports made available, 115 of these received no comments and therefore have a rating of zero. The 91 reports which remain and which are shown in Figure 91 come from 47 different matches within the corpus[10]. The distribution of rated reports against matches has a mean $\bar{x} = 2.215$ and a variance $\mu = 1.280$.

One other approach which we could have taken in deciding the importance of match reports was to use the television viewing figures for each match as a gauge of public interest. One could use the attendance figure for each game also, although this is not appropriate; the attendance is limited purely by stadium capacity and not by the level of interest within the game. The distribution of Irish viewing figures[11] for each of the games of the FIFA World Cup 2006 against comments per game in the SportsAnno corpus are shown in Figure 92. There is a positive correlation of 0.433 between the two sets of figures. We can see that there are a few outliers towards the top of the annotations count, as well as a grouping of matches with no annotations. One of the main reasons for a fall in correlation we feel may be the presence of these un-annotated games on the left of the graph.

[10] 54 games of the World Cup were recorded and with each game's video we had presented 3 different sources from different newspapers. For more details, see Page 61.

[11] These figures are not publicly available and have been supplied by the national Irish broadcaster, RTÉ. Figures were originally collected by *AGB Nielsen Media Research*.

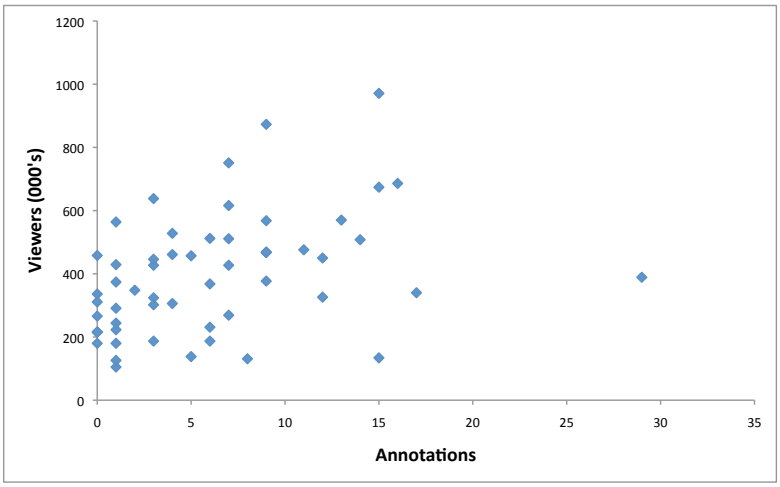


Figure 92: Scatterplot showing the correlation of annotation threads to RTÉ viewing figures (in 000s) for games during the FIFA World Cup 2006.

6.8.2 Searching Against the SportsAnno Corpus

As we have done with the SIGIR corpus, we created two indexes using the Lemur Toolkit against which we perform our searches. These two indexes are constructed in the same way as the SIGIR indexes of the last section. The first index, Sports_Txt contains documents made up of the original text from within each report. We include all 246 reports, in order to see the effect of comment-text inclusion more clearly. The second index, Sports_Comb, consists of documents made up of the original report text, but this time augmented by any comment-text made on that report. That is, all comments contained within a thread attached to the report. As we have said, unlike the case of SIGIR, this text is new to the index, and does not exist within any other document. For this reason, the index created in Sports_Comb is larger than that of Sports_Txt.

In the case of SIGIR, we had selected our search topics by using the section headings from within SIGIR's own proceedings. Since there is no direct analogy to these headings within the FIFA World Cup, we have chosen those 'topics' which best represent the competition as a whole. We have then augmented this set with a few topics which are more specific to our user community.

Table 28 shows the 9 queries which we have issued against the two indexes containing the SportsAnno comments. 3 of these queries represent awards granted as part of the competition itself. They have been selected as these terms appear to be of general interest to any person who would have followed the FIFA World Cup. The inclusion of the query "Cannavaro" is based on the performance of the player throughout the tournament; the Italian captain was also awarded the Golden Ball by UEFA as the European Player of 2006 [13], as well as FIFA Footballer of the Year[14]. Zidane was also

[13] http://www.uefa.com/competitions/ucl/news/kind=1/newsid=484425.html
[14] http://www.fifa.com/classicfootball/awards/playeroftheyear/winnermen.html

Table 28: Query terms issued against the SportsAnno indexes. All competition information and facts taken from the official FIFA website[12].

Query	*Description*
"Italy"	Winners of the tournament, after a faltering start early on.
"England"	The focus of much attention within the SportsAnno users community probably due to the lack of national team presence
"Zidane"	Winner of the Golden Ball as best player of the tournament, a highly contentious decision after head-butting an opposition player
"Cannavaro"	Many people's pick for Golden Ball winner due to great performances throughout the competition. Also the bookie's favourite
"Goal"	Included due to the fact that it is the most important word in the vocabulary of football
"Argentina"	Scorers of the competition's best team goal and also the best individual goal
"Klose"	Winner of the Golden Boot as top scorer in the competition,
"Australia"	Again the focus of much attention within the SportsAnno user community, gaining support in place of the absent national team
"Henry"	A footballer who had been present in the highly followed English Premiership, and had been linked with transfer talk during the year

the topic of much debate after head-butting an opponent in the chest during the World Cup final. It was his last game before retirement.

The remaining three queries are included as they represent the greatest interest which the SportsAnno community had within the tournament. The opening England vs. Paraguay game received significantly more comments than any other game in the competition. As described in Table 28, both Australia and Henry were of interest due to the absence of the Irish national side in the competition, and because of rumours which were present at time of the competition[15].

Using the 9 queries above, we perform Boolean retrieval and take the top-ranked 10 documents as rated by our group of users. We also give more weight to documents in which the query term appears in the title of the document. We do this by multiplying the rating given to the document by the number of comment threads created in the document. As we have shown, the number of annotations on a document is correlated to the general interest in terms of viewing figures. An example of the documents returned for the query "Italy" may be seen in Table 29. These 10 documents and the ranking order they appear in make up the ground-truth against which we shall compare the performance of our algorithms. The top 10 documents were chosen both to replicate the situation of the SIGIR corpus, and as a result of the work of Silverstein et al. (1999)

[15] http://english.people.com.cn/200605/22/eng20060522_267460.html

Table 29: Ranking for reports returned in reply to the query 'Italy'[16].

Rank	*Game*	Viewers	*Source*	*Threads*	*Comments*	*Rating*
1	Italy - Germany	BBC	873	9	5	2.555
2	Italy - USA	BBC	450	10	4	3.111
3	Italy - France	BBC	971	10	5	2.304
4	Italy - USA	Guardian	450	4	3	3.375
5	Italy - Australia	Guardian	326	7	3	3.333
6	Italy - France	Guardian	971	3	2	3.375
7	Italy - Ukraine	BBC	427	2	2	3.000
8	Italy - Australia	BBC	326	4	2	3.000
9	Italy - France	Sky	971	2	2	2.833
10	Italy - Ukraine	Guardian	427	1	1	3.375

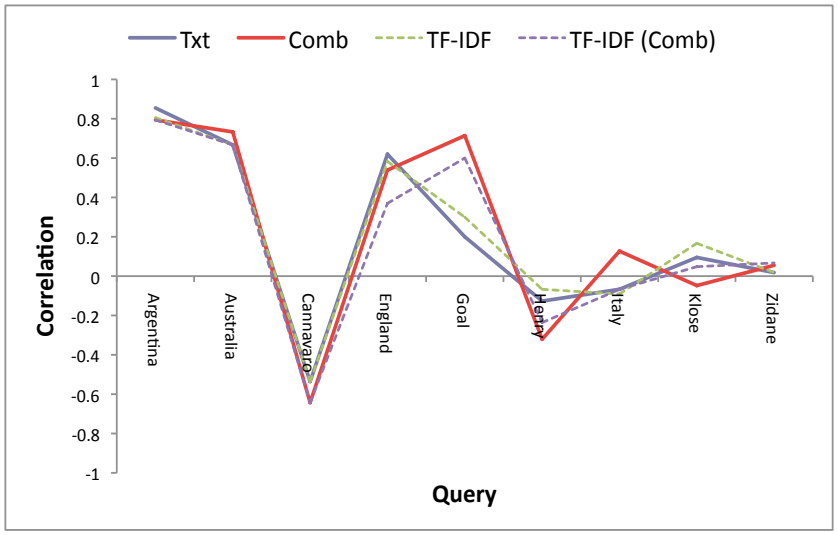


Figure 93: A comparison of the correlations achieved by TF-IDF and the MessageRank/AuthorRank algorithm against the rankings created by the aggregation of comment ratings on the Sports_Txt(Txt) and Sports_Comb(Comb) corpora. $A_R M_R$ uses the weights trained on the SIGIR corpus in the last section.

who noted that searchers are rarely interested in results outside of the top 5-10.

6.8.3 Using A_r and M_r to Re-Rank

The performance of AuthorRank and MessageRank on the SportsAnno corpus allows us to see how well the weights which we have trained on the SIGIR corpus can be transferred to a second smaller, real-world data-set. Figure 93 shows the comparison of the TF-IDF baseline correlation, and that of the ranking produced by Equation (35) (see Page 189) using the weights from the previous section. We can see that the performance of the algorithm is mixed with both higher and lower correlation with the ratings-based ground-truth. The differences between the TF-IDF and $A_R M_R$ rankings

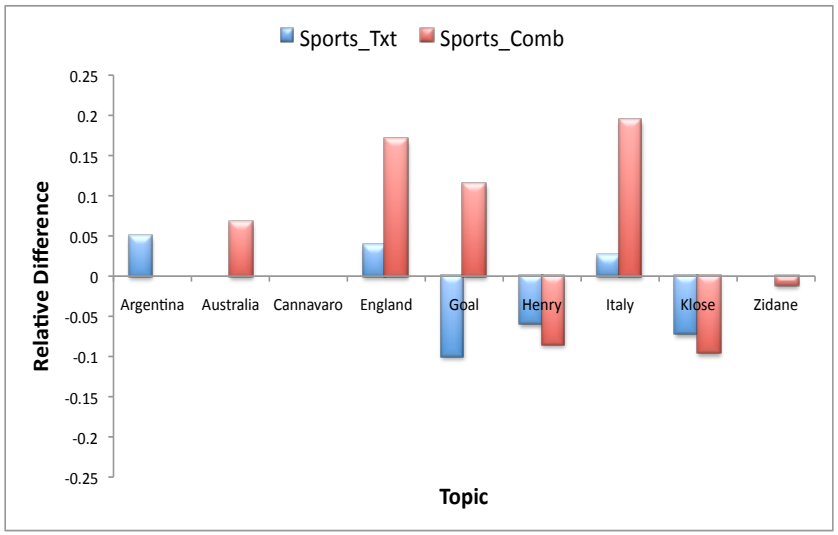


Figure 94: The change in correlations achieved by the MessageRank/AuthorRank algorithm compared to that of the TF-IDF score on the Sports_Txt and Sports_Comb indexes respectively. The performance of the algorithm is compared to that of TF-IDF on the same index.

are not significant on either the Sports_Txt ($p = 0.228$) or Sports_Comb ($p = 0.147$) index.

We can see from Figure 94 that in the case of specific player names (Cannavaro, Klose, Zidane, and Henry), there is a significant drop in correlation ($p = 0.092$) between the TF-IDF measure and that of the AuthorRank/MessageRank combination. We can observe in Figure 93 that in these cases, the correlation achieved by TF-IDF with the ratings based ground-truth also drops. This may be due to the weighting of documents containing the query when creating the original ranking. The names of the specific players rarely appears in the title of the reports, and in the cases of "Klose" and "Cannavaro", never. There is however a significant increase in correlation with the ratings ground-truth in the case of the queries 'England', 'Australia' and 'Italy' ($p = 0.034$) on the Sports_Comb index, as well as a reverse in the relative correlation for the query 'goal'. These improvements highlight the interests of the community of SportsAnno users as stated when choosing the query topics. The increases in 'Australia' and 'England' may be explained by the descriptions within Table 28, while the increases in the other topics may give a general indication of the interesting events within the tournament/corpus. Italy were the eventual winners of the tournament, while goals are of obvious interest to the community in general.

While our results were obtained through investigation of a relatively small corpus, they do indicate that the use of author and comment features in the ranking and re-ranking of query result sets is a useful direction of study. Using these features we have shown that it is possible to mimic the behaviour of a searcher with more expertise in a field. The benefit of this is that we are now able to return a ranking of documents

based not only on the content of the documents, but also on the information which can be gleaned from the social network of users who interact with the documents and each other. Using these features we are able to provide a browsing and search experience which is more social, allowing the user community to benefit and learn from each others' actions. It would also appear from the results in Figure 94 that the measures introduce a means of perceiving what the community found of interest as opposed to what is simply 'relevant'. Using author and message features appears to be a viable way in which to help users become part of the community, helping them understand the conversations and view-points of those around them. In doing so it is hoped that a more interactive and enjoyable online experience may be made possible.

CHAPTER VII

CONCLUSIONS AND SUMMARY

In this chapter we outline the conclusions which we have come to as a result of the work presented in this thesis. We shall re-visit the ideas and hypothesis put forward in Chapter 1, and comment on how the experiments we have carried out have performed in the context of these original plans. We look at our results and place them in the context of the research questions we had set out to answer when beginning our research. Finally we discuss some of the limitations of the experiments and data-sets we have used, before presenting some ideas on future work and directions this research might take.

7.1 Hypothesis Re-visited

In introducing the work we have done in this thesis, we stated an original hypothesis. This hypothesis aimed to encapsulate the idea that the social activities of recommendation and conversation could be used in an online environment to improve the quality and enjoyment of the online experience for a user. Our hypothesis was:

> *"The ranking of documents returned in answer to a user's information need may be improved by incorporating information from the social network of a documents' authors, as well as the network of annotations on the documents themselves."*

In order to investigate and prove the validity of this hypothesis, we have grounded our work in the fields of trust, social network analysis, and data-quality. We first presented two studies that we carried out into the usage patterns of two Web 2.0 systems designed to allow the functionality we state to be of use. These systems allowed for the annotation and viewing of currently disparate sources and mediums of sports presentation. We implemented these systems to allow their users to create in-context discussion threads, while simultaneously presenting corroborating evidence to any points they might make. The aim here was to help in the generation and continuation of discussion within the user community.

After discussing the outcome of these experiments, we extended our observations to a new and larger set of pseudo-annotations. This set was made up of citations on the SIGIR proceedings from 1997-2007. Before collecting the data-set, we established the connection and parallels between annotation and citation allowing us to move from one to the other. Once there we performed extensive experiments on our extended SIGIR data-set, in order to establish what features of users' annotation and citation behaviour are of most use in aiding the retrieval process. Finally, we presented the outcome of the experiments; we ascertained the effectiveness and promise of our algorithms to take advantage of the social and annotation networks of users when performing social information retrieval.

7.2 Research Objectives Re-visited

In order to test the hypothesis presented in Chapter 1, we identified a number of research questions which we believed would lead to the establishment of our hypothesis. We now iterate through these questions and highlight to what level we feel our research has answered them.

7.2.1 Annotation

We first look at the questions which concern the actual creation of annotations within a document corpus, before discussing the power of these annotations.

1. If users are given the opportunity to annotate documents, will they do so?
    i) Do users find the annotations of others within the community interesting?
    ii) Do users enjoy the additional interaction and social element which is introduced through the use of annotation?
    iii) Do users value the contribution of others?

We feel that the experiments presented in Chapter 3 verify past assertions of the importance and value of annotations as well as our own (Golovchinsky et al., 1999; Shipman et al., 2003; Marshall, 1997). Looking at the results of these experiments we can see that when given the opportunity to create annotations, users do so in order to engage in conversation. We can see that it is not just to create comments of their own, but also to reply to what others have all ready said. As a result of this, we believe that question i) is satisfied and that users do find the annotations of others of interest.

The results of both the informal survey conducted after the completion of the SportsAnno experiments, as well as the Annoby questionnaire show that users do enjoy the

opportunity to interact with each other. In Chapter 6 we have also shown the ratings which were given to the comments created within the SportsAnno corpus. These ratings show that the annotations created by other users were of interest to others within the community, another reason to incorporate these annotations into the ranking scheme of a search system. These ratings, along with the views expressed by users lead us to believe that the answer to both question ii) and iii) is yes; the community values both the opportunity to create annotations on a corpus, and the annotations which it creates.

7.2.2 Utility

After completing our system experiments, we evaluated the possibility of developing algorithms which could properly take advantage of annotations that the community has found both interesting and valuable. We introduced AuthorRank, (A_R), and MessageRank (M_R) which have been developed for this purpose. These algorithms aim to utilise the social and annotation networks of the user community to answer our second set of questions:

2. Are the annotations that users create on a 'social web' corpus of use to the user community as a whole?
    i) Can these annotations be leveraged to improve the overall performance of the system in satisfying users' information needs?
    ii) Can we identify specific elements of a user's profile of interactions which are of use in the ordering and ranking of documents to benefit the user?
    iii) Can the processes of "word-of-mouth" and "voting with your feet" be automated?

After establishing the strong similarities and comparability of the annotation and citation processes, we detailed our own collection of the citation and author network of SIGIR proceedings. We have shown in Chapter 5 that the way in which current citation search-engines rank and retrieve appears to be in a 'lowest-common-denominator' fashion. Our collection of an expert ground-truth to rank cited documents from within our SIGIR corpus against leads us to believe that the annotations/citations that users/authors create can be of use in improving the ranking of documents in answer to an information need. Annotations/citations help to provide an insight into the expertise of the authors creating them; this insight may then be used to improve the ranking algorithms, more-closely emulating the decision-making process of experts with even higher expertise.

In order to discover those elements of a user's profile that can be of most use in re-ranking retrieved documents, we examined the ability of each of our chosen features

to re-rank a TF-IDF baseline and improve correlation with an expert ground-truth. In doing so we have answered ii), discovering those features of the author and citation network that are most powerful and therefore of use in improving our ranking of retrieved documents.

We then attempted to establish answers to question i) and iii), combining the features using a variety of weights to leverage the strengths of each. In doing so we have shown that the ability of our algorithms to utilise the citation and author network in improving the rankings of retrieved documents is significant on our SIGIR indexes. Moving across to the original SportsAnno corpora we see a loss of significance. We shall discuss this, and our conclusions in the following section. In the context of our extended SIGIR corpus however, we feel that we have shown evidence to support our hypothesis as well as answer the questions we posed in Chapter 1.

7.3 Conclusions

In the sub-sections below we will outline our individual conclusions, based on the empirical studies carried out in this thesis. We will then draw some conclusions based on a user study with the prototype implementation of the proposed system.

7.3.1 Annotation Creation

When given the opportunity to create annotations, users will take advantage of them with the twin aims of allowing others to see their own points of view, and staying aware of what others are saying. Annotations proved a welcome addition to the Annoby and SportsAnno system, generating conversation and interest. The opportunity to annotate across different media representations of the same events (as was the case with Annoby) did not seem to have an effect on the number of annotations created. This may also be a result of the genre of the content; sports reports are designed to be a written description of the associated sports match. In a different genre however, say medical, historic, or educational content this may not be the case. The addition of annotations to a visual medical record may be of substantial additional benefit to written document annotations.

When creating a system which allows users to create annotations, some important considerations should be made. A notification system must be in place that allows users to quickly return to points of interest, or to reply to any comments left for them. Integration of an instant messaging (*IM*) client may also prove useful. In context commenting does indeed produce a style of annotation that is focussed and direct, limited to the context of the annotation. This has been shown to be true in both of our annotation

systems, as well as in past research by Marshall (1997).

While our systems were designed to allow anyone to partake in a community conversation, our experiments revealed that users prefer to comment on events they have already watched, rather than using the system to come up to speed with topics of conversation. Having said that, the second most common reason for using the systems behind the creation of annotations, was to watch the highlights of live matches that had been missed. We therefore feel that, although it was not evidenced by users viewing highlights they had not seen before, the system did allow users to reacquaint themselves with events of interest before beginning to annotate.

7.3.2 Expert Opinion

The additional considerations that more knowledgeable assessors give to creating a ranking of documents are significant, creating a ranking which is wholly different to that of less expert assessors. We have shown that these considerations are built on external knowledge not present within the documents themselves. Instead they come from knowledge of the meta-data associated with documents. In the case of our SIGIR proceedings this meta-data includes things such as institutional and author reputation, citation history, and semantic features like the scope of the document's content, structure etc. Of these, author reputation and citation history are something we have attempted to incorporate through the use of our chosen features. Using the features provides additional context and external information akin to that used by our experts to help in deciding their ranking of superior documents.

As stated before, while it may be argued that the ranking provided by services like Google Scholar (GS) are designed to best fit users' expectations and therefore needs, we do not feel that this ranking is the optimal ranking. The incorporation of features that provide some of the additional knowledge akin to that used by experts in creating a ranking should, by its nature, aid in simulating their style of ranking. We aim to provide a ranking of documents which is more closely aligned to that of an expert user, rather than simply one most anticipated by a more novice user.

We do make the observation that our expert set is quite small, although we have taken measures to ensure the rankings provided by our experts are statistically equivalent. While the addition of extra ratings may change our topics' final expertise scores, the ratings which have been gathered show that there is a negative correlation between the ranking scheme of Google Scholar, and the expertise of our raters. More extensive experimentation is warranted to see whether this fact remains true as the number of raters increases.

7.3.3 Author Network Features

Author information did prove useful in the case of SIGIR in increasing correlation with the expert ground truth of Chapter 5. Although many of the features that we looked at singly were not able to provide a significant increase in correlation, their combination did. Features based on the word count of authors, as in their expressiveness per post, were not of significant value to our calculations. We do however caution discrediting these features completely, since the nature of the data-set we used was such that significant variations in posting/message length were not possible. Consequently, every author would have had an average, and similar score with regards to word features. That is with the exception of overall total word count, since some authors wrote/cited more than others.

Of the features that we examined, we found that those features that are in some way network based, such as the 'average responses', 'started', and 'replied', were of greatest benefit in achieving significant correlation increases. Work on the SIGIR_Comb index (having included the citation-contexts within the bag-of-words of documents) did not see as many significant increases due to feature inclusion. This may be explained by the fact that the inclusion of the messages themselves is a more direct method of gauging the value of, say, a paper. Due to the length normalisation present in the Lemur Toolkit however, this can not be the only reason for increases in the correlation, leading us to believe that author features are indeed of note.

The performance of our AuthorRank algorithm was better than that of a straight linear combination, more than likely due to it's increased penalisation of 'barren' messages, whilst up-weighting responses. This result was encouraging all the same, showing that the incorporation of author features are indeed not just of note; they are significant in helping to improve the ranking of retrieved documents for a query bringing them more in line with that of an experts' ranking.

7.3.4 Citation/Annotation Network

The incorporation of annotations in the form of citation-context within documents results in a boost in correlation with our expert rankings. Taking into account not just the text of these annotations, but also the source of these annotations proved to also be of use (Larsen and Ingwersen, 2006, 2002). We have looked at the citation network as a source of additional information about a document's citations, and by inference the document itself. In the case of messages, the vocabulary-based features such as 'thread words' and 'average thread words' did prove of significant benefit. Again it should be noted that the usefulness of the 'average thread words' in comparison to that of 'message words' should not be over-estimated. It is however an interesting result regardless of

the near-uniformity of the 'message word [length]' feature, especially in cases such as micro-blogging as will be discussed in Section 7.5.

Of the features we have examined, the most effective in improving correlation with our experts' ranking when combined with TF-IDF seem to be those features that incorporate contextual information about the message, rather than relying on just the message itself. The boost provided by any single feature is not as significant as that provided by the h-slash re-ranking. Combination of these features again provides a 'sum-of-its-parts' outcome, as the performance of MessageRank is significantly better than that of any single feature. It is unable to significantly out-perform a linear combination of all features, but that in itself is not a problem since these same features are the features we have set out to show are of significant benefit in emulating an experienced users' ranking choices. Most interestingly, although the best single features were the network-based message features, the optimal combination of all features sees only 'message depth' given any weight. This weight is 50% of the combinations total though, showing network-based features are of use.

7.4 Considerations

We have already noted some of the considerations that should be taken into account due to the nature of the corpora used over the course of our experiments. Before beginning our experiments, we have shown that the theoretic basis for using citations in lieu of a large annotation corpus is sound. There are many consistencies in the method and reasons of use for each. In doing so however, the exact nature of the corpus has meant that certain characteristics could not be avoided.

The length of citation-contexts was chosen heuristically, and subsequently validated by the work of Ritchie et al. (2008). This choice did create a constraint on the variance within the corpus of the citation length. In a more real-world scenario we would expect to see far more variation in message length, as was the case in the SportsAnno and Annoby corpora. For this reason, we have advised against the complete disregard of word-based features of either author or message.

The choice of conference proceedings has enabled us to create a set of high-quality documents, meaning that the rankings created for each topic contain high-quality documents. This choice also assures the citation of adequate numbers of the documents, since the quality of the papers ensures citation. Proceedings did introduce the skewing of replies to new threads, since only a select number of our papers could ever be ensured to be new. The nature of research means that many of the papers within a conference will reference papers from past proceedings. The act of doing so means that such a paper is considered a citation of an earlier paper, and not part of the 'started' feature-set of

an author. This is perhaps one reason why the significance of the results on the SIGIR data-set did not transfer to the SportsAnno data-set.

This statistical significance itself may be of a slightly questionable nature. Since several statistical tests were performed, it may have been appropriate to perform a Bonferroni correction on the data (Abdi and Salkind, 2007). The correction is applied to limit the number of Type-II errors (false-positives) during tests for statistical significance. If we wish to find statistically significant results at a p-value of α across a series of n hypothesis tests, the correction of α/n must be applied.

The test has sometimes been criticised for being overly conservative and as shown by Cabin and Mitchell (2000), the exact situations in which to perform these corrections is sometimes difficult to decide. For example, when stating that one combination of weights is statistically significantly better than the TF-IDF baseline; this significance might be compared using a p-value of not 0.10, but $0.10/n$. In the case of the single features, this n would be 20 leading to no significant increases against the baseline's performance.

The counter argument may also be defended that the comparisons of different weights against the baseline is not the same hypothesis, since each set of weights is not also being compared. We have reported the best performing set of weights in each of the experiments in Chapter 6. Bonferroni correction is commonly used in the situation where several hypothesised variables are being tested simultaneously against a null hypotheses that they are all the same. As Perneger (1998) also points out the test itself it "concerned with the wrong hypothesis" in so far as it allows us to know if a set of variables are indeed statistically different, but does not tell us which of these variables nor how many.

The ground-truth created on the SportsAnno corpus of user ratings is subtly different to that of the rankings for SIGIR topic queries. Since there are also 3 documents per match, it could be possible future work to combine together the reports into a single document. Annotations could then be combined also, but the majority of annotations occur on the first-seen (*BBC*) report (see page 61). As such this approach may not work or be suitable for this particular corpus. It was also necessary to project the ratings received by comments onto the reports, losing some of the differentiation between comments. Ideally a ground-truth of report ratings would be created, although again the presence of the 3 reports per game would present problems. All of these reports are professionally produced and taken from reputable sources. As such, there is not quite the variation of, say, a blog environment where publishing standards can vary.

A final important consideration that must be made is towards the manner the Experts were asked to rank the papers with which they were presented. The use of a full ordinal ranking as opposed to a rating scale presents problems when ascertaining the

effects of variance within our experiments. A standard technique of ANOVA is not possible due to the ordinal nature of our data, as well as it's non-parametric characteristics. The use of rankings as opposed to ratings however further rules out the application of non-parametric tests such as the Kruskal-Wallis analysis of variance. In re-running these experiments we believe that a more statistically sound and clearer picture may be obtained by asking the experts to perform their judgements not as rankings, but as ratings. While these judgements would remain ordinal in nature, many more tests may be performed (such as the Kruskal-Wallis analysis of variance) to show the effects of variance within the different variables of our experiments.

Due to the difficulties with performing an analysis of variance within our experiments, it is no longer possible to perform the G-Study described in Bodoff and Li (2007). This study, as application of generalisation theory, helps to show the effects a-priori of changes to the key variables within an experimental set-up. The study allows us to see what effect would be seen had we chosen to utilise our experts in a different fashion. It is an important consideration when coupled with the findings of Voorhees (2000) that "as few as 25 topics can be used to compare the relative effectiveness of different retrieval systems with great confidence" but a golden-stadard is achieved around 50 topics. The agreement within our expert judgements has however been shown to be consistent across the 12 topics which we have chosen using the Kendall's Coefficient of Concordance (as detailed in Appendix B). It would however be a possible direction of future work to explore, spreading the expert judgements more thinly across more topics. This would be necessary since as we have noted, the increased work-load per expert in performing twice as many topic rankings was undesirable.

7.5 Directions for Future Work

In this section we outline some possible directions for future work, as well as some areas of application that could benefit from the adaption and adoption of this work.

Relevance Feedback The information provided by the citation-context, especially the additional index terms which are created after the insertion of these citations, has been shown to be of benefit in increasing the baseline performance of the TF-IDF algorithms. Relevance feedback (Salton and Buckley, 1990; Harman, 1992) can allow users to iterate through search steps, fine tuning the inclusion of documents that they have judged to be relevant. This manual iteration may be automated through the use of *pseudo relevance judgements*; documents that are highly-ranked by a system are assumed to be relevant, and are therefore included into the calculations of the next iterative loop. Using terms found within these top documents as *query expansion* terms may provided a similar boost to that of the citation-contexts. Here though, the inclusion of terms is

not decided by the annotations written by any single author, but instead is based on the top-ranked documents.

Clustering Techniques Clustering techniques may help in identifying which authors are of most importance within the context of a particular search. This would allow for improvements that are query-specific, while utilising the query-independent measure that have already been calculated. Authors may be weighted in accordance with the number of annotations, documents, or links they have created within the top subset of returned documents. Kurland and Lee (2004, 2005) use this approach to improve the results achieved through language-modeling techniques. We see it as a means of using the social network of users to discover the expertise and interests of users. This may in turn feed into the approaches to personalisation discussed below. Larsen and Ingwersen (2002, 2006) uses the co-citation and occurrence of citations between scientific papers to cluster documents for ranking and retrieval. The work on the 'boomerang effect' is also of interest in expanding the set of potentially relevant documents for a search query.

Spread Maximisation The premise of spread maximisation (Domingos and Richardson, 2001) is to maximise the spread of information across a network. It is a commonly used approach in the fields of viral marketing, where the information is spread by the agents themselves. Much work has been done on discovering those agents with the highest value; that is the agents who can help to spread the information to as many people as possible (Even-Dar and Shapira, 2007; Kempe et al., 2003). Recognising these people in particular can lead to a greater gain in advertising and sales, while utilising less time and money. This discovery of influential agents within the network is akin to the work presented in this thesis. We would like to investigate the utility of the approaches presented here in the field of spread maximisation; are the authors who prove the most interesting within the community of annotator also those who can aid the spread of information through the network? A first guess would be that in many cases yes, but not always. In our work, we do not make a distinction between authors who provide quality comments, and those who create comments which incite others. The first group of these annotators, who create quality information on which others comment, are we presume the same ones who would be of market value.

Summarisation The power of annotations to aid in finding important and useful information within documents has been studied before (Shipman et al., 2003), though this research was carried out on physical annotations rather than digital. Delort (2006) used the comments in blogs to aid in finding useful information, although he noted that this is a difficult problem due in essence to the lack of immediate context. In this regard, the systems we have built have helped to alleviate this problem, providing

a means of contextualising every comment. In conjunction with the features we have studied, these comments are far more likely to provide a means of locating the salient points within an article. Summarisation (Luhn, 1958; Kupiec et al., 1995) would also benefit from this sort of approach, as a key into points-of-interest for the community also provides a possible map of those elements of an article that are of greatest value. This summary however may be different in nature to one produced using the current extraction techniques, since this summary is less based on textual characteristics of the documents, and more on the social interest it generates. Boydell and Smyth (2007) have looked at the application of social summaries in previous work.

Personalisation Using the summarisation and clustering technique discussed above, there is an opportunity to personalise the result set returned to users. At present we focus on the features of users in the context of the social network as a whole. In order to personalise the results to a user, more specific information on the neighbours of users within their social network may be gathered. Along with this information, it is possible to discover which users appear most often in the context of specific searches, or indeed specific topics. Smyth et al. (2004) have used a similar approach in utilising the search patterns of users to improve the rankings of results chosen by similar users.

Micro-Blogging Micro-blogging and real-time search have been receiving much interest in recent times from both main-stream media[1] and the research community (Java et al., 2007; Jansen et al., 2009; Honeycutt and Herring, 2009). The ability to search sites such as Twitter[2] for information provided by its users is of great interest, as it allows for faster propagation and utilisation of information that in some cases may be time-sensitive. In this case again we see that the discovery of more credible or interesting sources of information is vital. Features of a user's social network, as well as the redistribution and linking to the information they create, is of intrinsic interest. We believe that the application of the measures and features we have developed may prove useful in this context. In this case however, we would look to incorporate more accurate data on the time annotations/messages have been in the system allowing for the introduction of additional temporal features.

7.6 Summary

In this thesis we have researched a style of social information retrieval that utilises not only the social network of users, but also that of the user-generated content produced. We have highlighted the opportunities for creating ranking schemes which exploit the

[1] http://www.aroundtheworldin140days.com/
[2] http://www.Twitter.com

social aspects of internet usage, while still providing a significant improvement on our baseline performance. This approach aims to show that the incorporation of social information can lead to a more enjoyable and useful user experience.

The continued growth and popularity of user-generated content along with technologies which aid in its generation and proliferation show there is a need for techniques that can take advantage of these new media. The way in which information is being produced for mass consumption is changing. We believe that the results shown here prove that this new media can be of use in both highlighting the valuable portions of a traditional sources of information, such as a newspaper article, as well as in its own right. User-generated content may be used to show what is of greatest interest to the community, as opposed to simply what may see most relevant. It remains an open research question as to whether these two things are one and the same, and is a question which we feel is deserving of further research.

APPENDIX A

ANNOBY USER QUESTIONNAIRE

Annoby Questionnaire

▶ **What did you use Annoby for most often? (You may choice more than one.)**

- ❑ Watch the highlights after watching on game live on TV
- ❑ Make comments after watching on game live on TV
- ❑ Watch when missed it the game live on TV
- ❑ In order to catch up what people think about the game
- ❑ Other: (please specify) _____

▶ **How many games have you watched on TV?**

- ❑ 0
- ❑ Around 1-5 games
- ❑ Around 6-10 games
- ❑ Around 11-15 games
- ❑ Around 16-20 games
- ❑ Around 21-25 games
- ❑ Around 26-30 games
- ❑ Most of the games
- ❑ All games

▶ **How frequently did you use the system?**

- ❑ Daily
- ❑ A few times a week
- ❑ A few times a month
- ❑ Never

Why?

▶ **How long did you spend using the system each session?**

- ❑ Usually less that 5 minutes
- ❑ Usually 5-15 minutes
- ❑ Usually 15-30 minutes
- ❑ Other: Please Specify

▶ Do you follow sports regularly? If so, do you normally do this online or using a different method? Why?

▶ Do you currently use any blogging software or actively participate in forum discussions on websites? If so, what blog/forum?

▶ Please rate (i.e., check an appropriate box) agreement or disagreement with the following statements regarding Annoby.

	Strongly agree	Quite agree	A little agree	Neutral	A little disagree	Quite disagree	Strongly disagree
The system is easy to use							
It is easy to make and read comments							
It is easy to understand people's comments and follow the meaning of threads							
It is useful to be able to comment on the video directly							
Reading what other people have to say is of interest to me							
I like to reply on other people's comments							
I like to comment directly on the report							
The system is fun to use							
The system allows me to be sociable with other users							
It is useful to be able to comment on specific text within the report rather than on the report as a whole							

▶ What were your favourite features of the Annoby system?

▶ What features would you have liked to see within the Annoby system?

▶ Any other comments you want to tell us?

Thank you very much!

APPENDIX B

KENDALL'S CO-EFFICIENT OF CONCORDANCE

Kendall's W measure is a non-parametric statistical test for the agreement amongst testers (e.g. experts asked to give a ranking of wines; a focus group asked to give a preference of political candidates; or in our case experts asked to order scientific papers by order of perceived usefulness). It is closely related to both Friedman's two-way analysis of variance without repeated ranks, and Spearman's ρ correlation coefficient. Kendall's W measures the actual amount of variation between judges' ranks against the expected variance as a consequence of chance. To do this we first compute the row-marginal sums of ranks R_i received by n objects. In our case, these objects are the SIGIR papers, and the rankings are provided by the p judges, or experts. This is then used to calculated the sum-of-squares statistic, S, over ranks R_i:

$$S = \sum_{i=1}^{n} (R_i - \bar{R})^2 \tag{36}$$

$\bar{R}$ is the mean of the R_i values, and W may now be calculated as follows:

$$W = \frac{12S}{p^2(n^3 - n) - pT} \tag{37}$$

where T is the correction for tied-ranks (in our case this may be ignored):

$$T = \sum_{x=1}^{m} ({t_x}^3 - t_x) \tag{38}$$

where t_x is the number of tied ranks in each (x) of m groups of ties. The sum is then computed over each of the p judges. As stated, W is strongly related to Spearman's ρ which gives the correlation between two judge's rankings (Siegel and Castellan, 1956). Kendall's W is in fact calculable from the mean, $\bar{r}$, of all the pair-wise Spearman correlations ρ using the formula:

$$W = \frac{(p-1)\bar{r} - 1}{p} \tag{39}$$

When testing for a statistically significant level of agreement amongst the ratings of judges, we first assume that there is a disagreement in the rankings.

H_0: *There is disagreement between the ratings of the judges*
H_1: *There is agreement between the ratings of the judges*

While W is a non-parametric measure, it may be used to closely approximate the χ_2 distribution:

$$\chi^{2(n-1)} = p(n-1)W \tag{40}$$

As $n \to \infty$, W provides a closer approximation of the χ^2 distribution with $n-1$ degrees of freedom (see Figure 95). Table 30 shows the W, $\chi^{2(n-1)}$ and p-values for each of the topics rated by our experts. All of the rankings are shown to have a statistically significant level of agreement, except for the *"cross-lingual"* topic. For this reason, we have chosen to ignore the topic from this point on due to the inconsistency of ground-truth measurements.

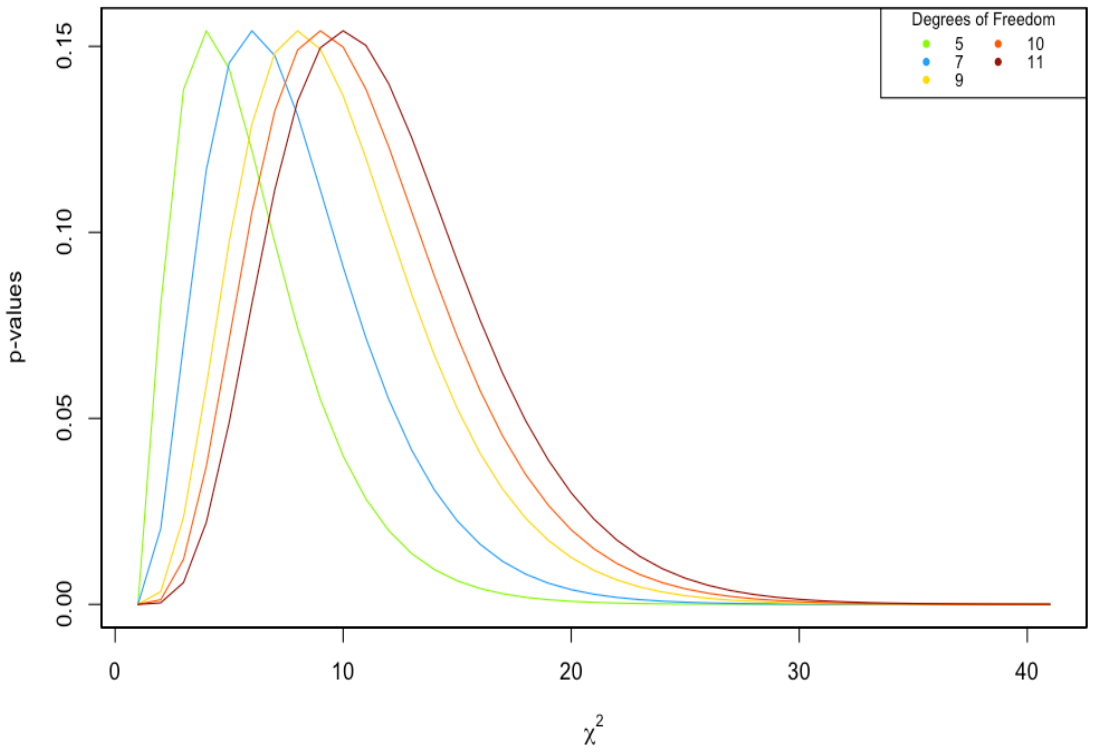


Figure 95: χ^2 distributions used in the expert rank comparisons

Table 30: Kendall's W and significance levels for per-topic inter-expert ranking agreement

Topic	*Papers (n)*	*Experts (k)*	*Kendall's W*	$\chi^{2(n-l)}$	*p-value*
Collaborative Filtering	10	7	0.558	35.2	$5.57e^{-5}$
Cross-Lingual IR	10	7	0.253	15.9	0.0683
Distributed IR	8	7	0.541	26.5	0.000408
Document Clustering	10	7	0.414	26.1	0.00199
Image Retrieval	11	9	0.308	27.7	0.00199
Language Modeling	12	8	0.5	44	$7.16e^{-6}$
Latent Semantic Indexing/Analysis	12	6	0.398	26.3	0.00594
Linkage Analysis	10	6	0.441	23.8	0.0046
Personalisation	*10*	*10*	*0.601*	*54.1*	$1.83e^{-8}$
Question Answering	9	7	0.335	18.7	0.0163
Relevance Feedback	10	8	0.525	37.8	$1.89e^{-5}$
Spam	6	7	0.380	13.3	0.0208
Text Summarisation	9	8	0.558	35.7	$1.96e^{-5}$
Topic Distillation	8	7	0.524	25.7	0.000578

APPENDIX C

XML AND MPEG-7

Extensible Mark-up Language[1] (XML) is a World Wide Web Consortium (W3C) recommended standard for the sharing and creation of information. It is becoming increasingly popular with web-publishers due to its extensibility; the format allows for the creation of user-defined elements within a document, similar to HTML but without the pre-defined naming conventions. Instead, every XML file follows an associated schema in which the type of information stored in any element is defined. It is one of the main technologies behind all social media shared across the web today. XML's extensibility means that users are not confined to learning a standard document model but may instead define their own schema for any desired task. We have used XML for storage of all the information about each SIGIR paper.

MPEG-7 The MPEG-7 standard, Multimedia Content Description Interface, defines the syntax and semantics of video descriptions (Manjunath et al., 2002). It was confirmed as an ISO standard in February 2002. Previous MPEG (Moving Pictures Expert Group) standards such as MPEG-1, MPEG-2 and MPEG-4 focussed on the encoding of the audio-visual signal; MPEG-7 does not specify any coded representation of audio-visual information but focuses on the standardisation of a common interface for describing multimedia materials. By using an XML base, the standard allows for an inter-operable description of the video which can be used by many different retrieval systems. It also provides a clean interface to individual video indexing tools which can be viewed as functional black boxes that take as input the video and its initial MPEG-7 descriptions, and outputs an updated MPEG-7 description.

The MPEG-7 Multimedia Description Schemes (MDS) provide general descriptions for content, its management, organisation, navigation, access and also user interaction (see Figure 96). The MDS allows content to be decomposed both temporally and spatially, thereby allowing description of sub-units such as shots, objects or regions. The MPEG-7 System tools provide a mechanism for the MPEG-7 standard, which is XML based, to be encoded in a compact binary representation and supports multiplexing and synchronising the description with the video content.

[1] http://www.w3.org/XML/

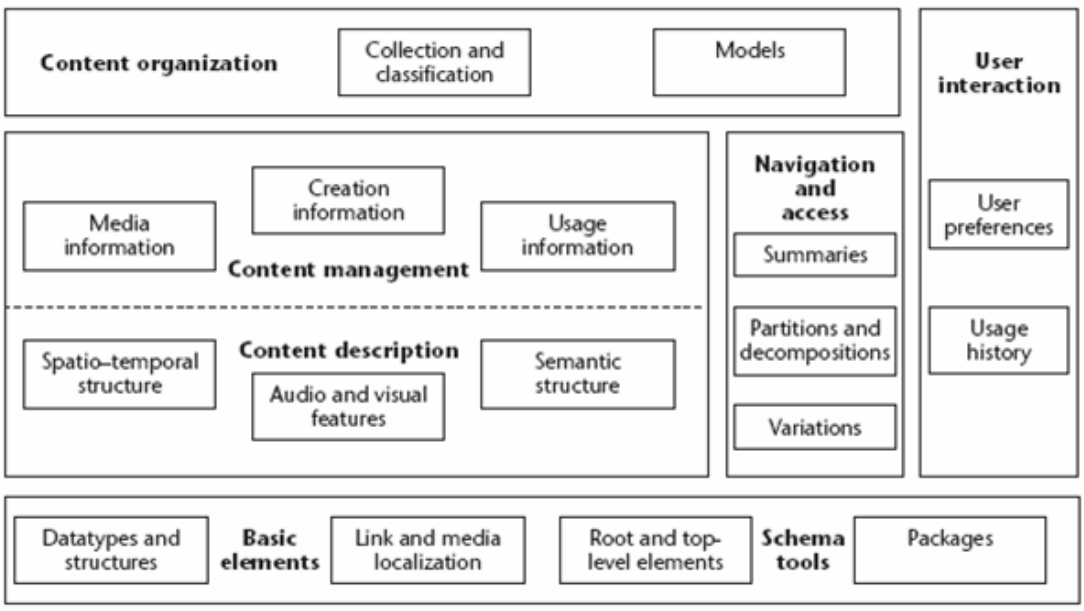


Figure 96: Overview of the MPEG-7 Multimedia description schemes

References

Abdi, H. and Salkind, N. (2007), The Bonferonni and Šidák Corrections for Multiple Comparisons, Sage Pub., Thousand Oaks, CA, USA.

Abdul-Rahman, A. (1997), 'The PGP Trust Model', The Journal of Electronic Commerce **10**, 27–31.

Abdul-Rahman, A. and Hailes, S. (2000), Supporting Trust in Virtual Communities, in 'Proceedings of the 33rd Hawaii International Conference on System Sciences', Vol. 6, Maui, Hawaii, USA, p. 9.

Abel, F., Frank, M., Henze, N., Krause, D., Plappert, D. and Siehndel, P. (2007), Groupme! - where semantic web meets web 2.0, in 'The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007', Springer-Verlag New York Inc, Busan, Korea, p. 871.

Adler, B. and de Alfaro, L. (2007), A Content-Driven Reputation System for the Wikipedia, in 'WWW '07: Proceedings of the 16th International Conference on World Wide Web', Banff, Alberta, Canada, pp. 8–12.

Adler, M. (1972), How to Read a Book, Touchstone.

Agosti, M., Bonfiglio-Dosio, G. and Ferro, N. (2007), 'An Historical and Contemporary Study On Annotations To Derive Key Features For Systems Design', Int J Digit Libr **8**(1), 1–19.

Agosti, M. and Ferro, N. (2003), Annotations: Enriching a Digital Library, in 'Research and Advanced Technology for Digital Libraries', Proceedings of the 7th European Conference on Digital Libraries, August 17-22, Trondheim, Norway.

Akerlof, G. (1970), 'The Market for Lemons: Qualitative Uncertainty and the Market Mechanism', Quarterly Journal of Economics **84**(3), 488–500.

Allan, J., Callan, J., Collins-Thompson, K., Croft, W. B., Feng, F., Fisher, D., Lafferty, J., Larkey, L., Truong, T., Ogilvie, P. et al. (2003), 'The Lemur Toolkit for Language Modeling and Information Retrieval'.
URL: *http://lemur.wiki.sourceforge.net/*

Andrews, P. (2005), Sports Journalism: A Practical Guide, Sage, chapter 5, pp. 48–50.

Arrow, K. (1962), 'The Economic Implications of Learning by Doing', The Review of Economic Studies **29**(3), 155–173.

Aurnhammer, M., Hanappe, P. and Steels, L. (2006), Augmenting Navigation for Collaborative Tagging with Emergent Semantics, in 'ISWC 2006: The 5th International Semantic Web Conference', Vol. 4273, Springer, Athens, Georgia, USA, p. 58.

Axelrod, R. (1984), The Evolution of Cooperation, Basic Books.

Baeza-Yates, R., Ribeiro-Neto, B. et al. (1999), Modern Information Retrieval, Addison-Wesley Harlow, England, pp. 193–194.

Balabanović, M. and Shoham, Y. (1997), 'Fab: Content-based, Collaborative Recommendation', Communications of the ACM **40**(3), 66–72.

Banks, M. (2006), 'An Extension of the Hirsch Index: Indexing Scientific Topics and Compounds', Scientometrics **69**(1), 161–168.

Bartell, B., Cottrell, G. and Belew, R. (1994), Automatic Combination of Multiple Ranked Retrieval Systems, in 'Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', Dublin, Ireland, pp. 173–181.

Bateman, S., Farzan, R., Brusilovsky, P. and McCallan, G. (2006), OATS: The Open Annotation and Tagging System, in 'I2LOR '06: Proceedings of 3rd Annual E-Learning Conference on Intelligent Interactive Learning Object Repositories', Montreal, Quebec, Canada.

Batini, C. and Scannapieco, M. (2006a), Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications), Springer-Verlag New York, Inc. Secaucus, NJ, USA.

Batini, C. and Scannapieco, M. (2006b), Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications), Springer-Verlag New York, Inc. Secaucus, NJ, USA, p. 38.

Batista, P., Campiteli, M. and Kinouchi, O. (2006), 'Is It Possible to Compare Researchers With Different Scientific Interests?', Scientometrics **68**(1), 179–189.

Begelman, G., Keller, P. and Smadja, F. (2006), 'Automated Tag Clustering: Improving Search and Exploration in The Tag Space', Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland .

Berners-Lee, T., Hendler, J., Lassila, O. et al. (2001), 'The Semantic Web', Scientific American **284**(5), 28–37.

Bharat, K. and Henzinger, M. (1998), Improved Algorithms for Topic Distillation in a Hyperlinked Environment, in 'Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', Melbourne, Australia, pp. 104–111.

Bodoff, D. and Li, P. (2007), Test theory for assessing IR test collections, in 'SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', ACM New York, NY, USA, Amsterdam, The Netherlands, pp. 367–374.

Bollacker, K., Lawrence, S. and Giles, C. (1998), CiteSeer: An Autonomous Web Agent For Automatic Retrieval and Identification Of Interesting Publications, in 'Proceedings of the Second International Conference on Autonomous Agents', Minneapolis, Minnesota, USA, pp. 116–123.

Bondy, A. J. and Murty, U. S. R. (1976), Graph Theory With Applications, North-Holland, New York.

Bornmann, L., Mutz, R. and Daniel, H. (2008), 'Are There Better Indices for Evaluation Purposes Than the h Index? A Comparison of Nine Different Variants of the h Index Using Data From Biomedicine', Journal of the American Society for Information Science and Technology **59**(5), 1–8.

Bottoni, P., Civica, R., Levialdi, S., Orso, L., Panizzi, E. and Trinchese, R. (2004), MADCOW: A Multimedia Digital Annotation System, in 'AVI '04: Proceedings of the Working Conference on Advanced Visual Interfaces', Gallipoli (Lecce), Italy.

Bottoni, P., Levialdi, S. and Rizzo, P. (2003), An Analysis and Case Study of Digital Annotation, in 'DNIS 2003: Databases in Networked Information Systems, Third International Workshop', Vol. 2822/2003, Aizu, Japan, pp. 216–230.

Boydell, O. and Smyth, B. (2007), From Social Bookmarking to Social Summarization: An Experiment in Community-Based Summary Generation, in 'IUI '07: Proceedings of the 12th International Conference on Intelligent User Interfaces', ACM, Honolulu, Hawaii, USA, pp. 42–51.

Briggs, P. and Smyth, B. (2007), Trusted Search Communities, in 'Proceedings of the 12th International Conference on Intelligent User Interfaces', Honolulu, Hawaii, USA, pp. 337–340.

Brooks, T. (1986), 'Evidence of Complex Citer Motivations', Journal of the American Society for Information Science **37**(1), 34–36.

Brush, A., Bargeron, D., Grudin, J., Borning, A. and Gupta, A. (2002), 'Supporting Interaction Outside of Class: Anchored Discussions vs. Discussion Boards', Proceedings of ACM CHI 2002, Minneapolis, Minnesota, April 20-25 .

Brush, A. J. B., Bargeron, D., Gupta, A. and Cadiz, J. J. (2001), Robust Annotation Positioning in Digital Documents, in 'CHI '01: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', ACM, Seattle, Washington, 31 March-5 April, pp. 285–292.

Buneman, P., Khanna, S. and Tan, W. (2001), Why and Where: A Characterization of Data Provenance, in 'Proceedings of the 8th International Conference on Database Theory', Springer-Verlag London, UK, pp. 316–330.

Burrell, Q. (2007), 'Hirsch's h-Index: A Stochastic Model', Journal of Informetrics **1**(1), 16–25.

Butler, D. (2004), 'Science Searches Shift Up a Gear as Google Starts Scholar Engine', Nature **432**(7016), 423.

Cabin, R. and Mitchell, R. (2000), 'To Bonferroni or not to Bonferroni: When and How Are The Questions', Bulletin of the Ecological Society of America pp. 246–248.

Cadiz, J., Gupta, A. and Grudin, J. (2000), Using Web Annotations for Asynchronous Collaboration Around Documents, in 'CSCW '00: Proceedings of the 2000 ACM Conference on Computer Supported Co-operative Work', Philadelphia, Pennsylvania, USA.

Cano, V. (1989), 'Citation Behavior: Classification, Utility, and Location', Journal of the American Society for Information Science **40**(4), 284–290.

Carter, J., Bitting, E. and Ghorbani, A. (2002), 'Reputation Formalization for an Information-Sharing Multi-Agent System', Computational Intelligence **18**(4), 515–534.

Chakrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Gibson, D. and Kleinberg, J. (1998), 'Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text', Computer Networks and ISDN Systems **30**(1-7), 65–74.

Chakrabarti, S., Joshi, M. and Tawde, V. (2001), Enhanced Topic Distillation Using Text, Markup Tags, and Hyperlinks, in 'Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', New Orleans, USA, pp. 208–216.

Chopra, K., Wallace, W., Aptima, I. and Woburn, M. (2003), Trust in Electronic Environments, in 'Proceedings of the 36th Annual Hawaii International Conference on System Sciences', Honalulu, Hawaii, USA, p. 10.

Claypool, M., Le, P., Wased, M. and Brown, D. (2001), Implicit Interest Indicators, in 'IUI '01: Proceedings of the 6th International Conference on Intelligent User Interfaces', ACM, Santa Fe, New Mexico, United States, pp. 33–40.

Cleverdon, C. (1967), The Cranfield Tests on Index Language Devices, Aslib Proceedings.

Cleverdon, C. (1972), 'On the Inverse Relationship of Recall and Precision.', Journal of Documentation **28**(3), 195–201.

Cleverdon, C. (1988), 'Optimizing Convenient Online Access to Bibliographic Databases', Taylor Graham Series In Foundations Of Information Science pp. 32–41.

Craig, E., ed. (2008), Routledge Encyclopedia of Philosophy, Routledge.

Croft, W. (2000), 'Combining Approaches to Information Retrieval', Advances in Information Retrieval **7**, 1–36.

Croft, W. and Harper, D. (1979), 'Using Probabilistic Models of Document Retrieval Without Relevance Information', Journal of Documentation **35**(4), 285–295.

Das-Gupta, P. (1988), 'Trust as a Commodity', Trust: Making and Breaking Cooperative Relations pp. 49–72.

Das-Gupta, P. and Katzer, J. (1983), A Study of the Overlap Among Document Representations, in 'Proceedings of the 6th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', Bethesda, Maryland, USA, pp. 106–114.

Dasu, T. and Johnson, T. (2003), Exploratory Data Mining and Data Cleaning, Wiley-Interscience.

Dellarocas, C. (2003), 'The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms', Management Science **49**(10).

Delort, J.-Y. (2006), Identifying Commented Passages of Documents Using Implicit Hyperlinks, in 'HYPERTEXT '06: Proceedings of the Seventeenth Conference on Hypertext and Hypermedia', ACM, New York, NY, USA, pp. 89–98.

Dey, A. (2001), 'Understanding and Using Context', Personal and Ubiquitous Computing **5**(1), 4–7.

Dodds, P., Muhamad, R. and Watts, D. (2003), 'An Experimental Study of Search in Global Social Networks', Science **301**(5634), 827–829.

Domingos, P. and Richardson, M. (2001), Mining the Network Value of Customers, in 'KDD '01: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM, San Francisco, California, USA, pp. 57–66.

Economist (2001), Keeping the Customer Satisfied. The Economist, July 2001.

Egghe, L. (2006), 'Theory and Practise of The g-Index', Scientometrics **69**(1), 131–152.

Erdös, P. and Rényi, A. (1959), 'On Random Graphs', Publicationes Mathematicae **6**(290–297).

Even-Dar, E. and Shapira, A. (2007), A Note on Maximizing the Spread of Influence in Social Networks, in 'WINE 2007: The 3rd International Workshop On Internet And Network Economics', Vol. 4858 of Lecture Notes in Computer Science, Springer-Verlag London, Springer, San Diego, California, USA, p. 281.

Fiore, A. T., Tiernan, S. L. and Smith, M. A. (2002), Observed Behavior and Perceived Value Of Authors In Usenet Newsgroups: Bridging The Gap, in 'CHI '02: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', ACM, New York, NY, USA, pp. 323–330.

Fox, E. A. and Shaw, J. A. (1995), Combination of Multiple Searches, in D. Harman, ed., 'Proceedings of TREC-3', NIST Special Publication, pp. 500–226.

Freyne, J., Farzan, R., Brusilovsky, P., Smyth, B. and Coyle, M. (2007), Collecting Community Wisdom: Integrating Social Search & Social Navigation, in 'Proceedings of the 12th International Conference on Intelligent User Interfaces', ACM Press New York, NY, USA, Orlando, Florida, USA, pp. 52–61.

Friedman, B., Khan Jr, P. and Howe, D. (2000), 'Trust Online', Communications of the ACM **43**(12), 34–40.

Friedman, E. and Resnick, P. (2001), 'The Social Cost of Cheap Pseudonyms', Journal of Economics & Management Strategy **10**(2), 173–199.

Frommholz, I., Brocks, H., Thiel, U., Neuhold, E., Iannone, L., Semeraro, G., Berardi, M. and Ceci, M. (2003), Document-Centered Collaboration for Scholars in the Humanities-The COLLATE System, in 'The 7th Annual European Conference on Digital Libraries, ECDL 2003', Vol. 2769/2003, Trondheim, Norway, pp. 434–445.

Garfield, E. (1965), 'Can Citation Indexing Be Automated?', Statistical Associtation Methods for Mechanized Documentation, Symposium Proceedings pp. 189–192.

Garfield, E. (1972), 'Citation Analysis as a Tool in Journal Evaluation', Science **178**, 471–479.

Garfield, E. (1997), 'Concept of Citation Indexing: A Unique and Innovative Tool For Navigating The Research Literature'.

Gilbert, E. and Karahalios, K. (2009), Predicting Tie Strength With Social Media, in 'CHI '09: Proceedings of the 27th International Conference on Human Factors in Computing Systems', ACM, Boston, MA, USA, pp. 211–220.

Giles, C., Bollacker, K. and Lawrence, S. (1998), CiteSeer: An Automatic Citation Indexing System, in 'Proceedings of the 3rd ACM Conference on Digital Libraries', Pittsburgh, Pennsylvania, USA, pp. 89–98.

Gladwell, M. (2000), The Tipping Point: How Little Things Can Make a Big Difference, Little, Brown and Company.

Godsil, C. and Royle, G. (2001), Algebraic Graph Theory, Springer.

Golbeck, J. and Hendler, J. (2004), Accuracy of Metrics for Inferring Trust and Reputation in Semantic Web-Based Social Networks, in 'The 14th International Conference on Engineering Knowledge in the Age of the Semantic Web, EKAW 2004', Vol. 3257/2004, Whittlebury Hall, Northamptonshire, UK, pp. 116–131.

Golbeck, J., Parsia, B. and Hendler, J. (2003), Trust Networks on the Semantic Web, in 'The 7th International Workshop on Cooperative Information Agents', Vol. 2782/2003, Helsinki, Finland, pp. 238–249.

Goldberg, D., Nichols, D., Oki, B. M. and Terry, D. (1992), 'Using Collaborative Filtering to Weave an Information Tapestry', Commun. ACM **35**(12), 61–70.

Golder, S. and Huberman, B. (2006), 'Usage Patterns of Collaborative Tagging Systems', Journal of Information Science **32**(2), 198–208.

Golovchinsky, G. (1997), What the Query Told the Link: The Integration of Hypertext and Information Retrieval, in 'HYPERTEXT '97: Proceedings of the 8th ACM Conference on Hypertext', ACM, Southampton, UK, pp. 67–74.

Golovchinsky, G., Price, M. N. and Schilit, B. N. (1999), From Reading to Retrieval: Free-Form Ink Annotations as Queries, in 'SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', Berkeley, California, USA, pp. 19–25.

Gómez, V., Kaltenbrunner, A. and López, V. (2008), Statistical Analysis of the Social Network and Discussion Threads in Slashdot, in 'WWW '08: Proceeding of the 17th International Conference on World Wide Web', ACM, New York, NY, USA, pp. 645–654.

Google Inc. (2006), 'Google Search Engine'.
URL: *http://www.google.com*

Granovetter, M. (1973), 'The Strength of Weak Ties', American Journal of Sociology **78**(6), 1360.

Guare, J. (1990), Six Degrees of Separation: A Play, Random House.

Guha, R., Kumar, R., Raghavan, P. and Tomkins, A. (2004), Propagation of Trust and Distrust, in 'Proceedings of the 13th International Conference on World Wide Web', New York, NY, USA, pp. 403–412.

Guo, L., Shao, F., Botev, C. and Shanmugasundaram, J. (2003), XRANK: Ranked Keyword Search Over XML Documents, in 'SIGMOD '03: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data', San Diego, California, USA.

Gyöngyi, Z., Garcia-Molina, H. and Pedersen, J. (2004), Combating Web Spam with TrustRank, in 'Proceedings of the 13th International Conference on Very Large Databases', Vol. 30, VLDB Endowment, pp. 576–587.

Halpin, H., Robu, V. and Shepherd, H. (2007), The Complex Dynamics Of Collaborative Tagging, in 'WWW '07: Proceedings of the 16th International Conference on World Wide Web', ACM, Banff, Alberta, Canada, pp. 211–220.

Harman, D. (1992), Relevance feedback revisited, in 'SIGIR '92: Proceedings of the 15nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', ACM, Copenhagen, Denmark, pp. 1–10.

Harman, D. (1993), Overview of the First TREC Conference, in 'Proceedings of the 16th Annual International ACM SIGIR conference on Research and Development in Information Retrieval', Pittsburgh, Pennsylvania, USA, pp. 36–47.

Harzing, A. and van der Wal, R. (2007), 'Google Scholar: The Democratization of Citation Analysis?', Ethics in Science and Environmental Politics **8**(1), 61–73.

Haveliwala, T. (2002), Topic-Sensitive Pagerank, in 'WWW '02: Proceedings of the 11th International World Wide Web Conference', Honolulu, Hawaii, USA, pp. 517–526.

Hearst, M. (1997), 'Text-Tiling: Segmenting Text Into Multi-Paragraph Subtopic Passages', Computational Linguistics **23**(1), 33–64.

Hermida, A. and Thurman, N. (2008), 'A Clash of Cultures: The Integration of User-Generated Content Within Professional Journalistic Frameworks at British Newspaper Websites', Journalism Practice **2**(3).

Hess, C. (2006), 'Trust-Based Recommendations for Publications - A Multi-Layer Network Approach', TCDL Bulletin **2**(2), 1–11.

Hiemstra, D., Hauff, C., Jong, F. and Kraaij, W. (2007), 'SIGIR's 30th Anniversary: An Analysis of Trends in IR Research and The Topology of Its Community', ACM SIGIR Forum **41**(2).

Hirsch, J. (2005), 'An Index to Quantify an Individual's Scientific Research Output', Proceedings of the National Academy of Sciences **102**(46), 16569–16572.

Hirschman, A. (1984), 'Against Parsimony: Three Easy Ways of Complicating Some Categories of Economic Discourse.', American Economic Review **74**(2), 89–96.

Hoffman, P. (1998), The Man who Loved Only Numbers: The Story of Paul Erdös and the Search for Mathematical Truth, Fourth Estate.

Honeycutt, C. and Herring, S. (2009), Beyond Microblogging: Conversation and Collaboration via Twitter, in 'HICSS'09: Proceedings of the 42nd Hawaii International Conference on System Sciences', Big Island, Hawaii, USA, pp. 1–10.

Hotho, A., Jaschke, R., Schmitz, C. and Stumme, G. (2006), Information Retrieval in Folksonomies: Search and Ranking, in 'ESWC '06: Proceedings of the 3rd Annual European Semantic Web Conference', Vol. 4011, Springer, Budva, Montenegro, p. 411.

Houser, D. and Wooders, J. (2006), 'Reputation in Auctions: Theory, and Evidence from eBay', Journal of Economics & Management Strategy **15**(2), 353–369.

Iglesias, J. and Pecharromán, C. (2007), 'Scaling the h-Index for Different Scientific ISI Fields', Scientometrics **73**(3), 303–320.

Jansen, B. J., Zhang, M., Sobel, K. and Chowdury, A. (2009), Micro-Blogging as Online Word of Mouth Branding, in 'CHI EA '09: Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems', ACM, Boston, MA, USA, pp. 3859–3864.

Jansen, B., Spink, A. and Saracevic, T. (2000), 'Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web', Information Processing and Management **36**(2), 207–227.

Java, A., Song, X., Finin, T. and Tseng, B. (2007), Why We Twitter: Understanding Microblogging Usage and Communities, in 'WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis', ACM, San Jose, California, pp. 56–65.

Jeong, H., Albert, R. and Barabasi, A. (1999), 'Diameter of the World-Wide Web', Nature(London) **401**(6749), 130–131.

Jeong, H., Neda, Z. and Barabasi, A. L. (2003), 'Measuring Preferential Attachment In Evolving Networks', Europhysics Letters **61**(4), 567–572.

Jin, B. (2006), 'H-index: An Evaluation Indicator Proposed by Scientist', Science Focus **1**(1), 8–9.

Joachims, T., Granka, L., Pan, B., Hembrooke, H. and Gay, G. (2005), Accurately Interpreting Clickthrough Data as Implicit Feedback, in 'SIGIR '05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', ACM, Salvador, Brazil, pp. 154–161.

Kahan, J. and Koivunen, M.-R. (2001), Annotea: An Open RDF Infrastructure for Shared Web Annotations, in 'WWW '01: Proceedings of the 10th International Conference on World Wide Web', Proceedings of the 10th International Conference on World Wide Web.

Kasturirangan, R. (1999), Multiple Scales in Small-World Networks, in 'Brain and Cognitive Science Department', Massachusetts Institute of Technology Cambridge, MA, USA.

Kempe, D., Kleinberg, J. and Tardos, E. (2003), Maximizing the Spread of Influence Through a Social Network, in 'KDD '03: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM, Washington, D.C., USA, pp. 137–146.

Kendall, M. and Smith, B. (1939), 'The Problem of m Rankings', Annals of Mathematical Statistics **10**(3), 275–287.

Kirsch, S. (2006), Social Information Retrieval, Master's thesis, Rheinischen Friedrich-Wilhelms-Universität, Bonn.

Kleinberg, J. M. (1998), Authoritative Sources In a Hyperlinked Environment, in 'SODA '98: Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms', Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, pp. 668–677.

Kleinfeld, J. (2002), 'The Small World Problem', Society **39**(2), 61–66.

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R. and Riedl, J. (1997), 'GroupLens: Applying Collaborative Filtering to Usenet News', Communications of the ACM **40**(3), 77–87.

Korte, C. and Milgram, S. (1970), 'Acquaintance Linking Between White and Negro Populations: Application of The Small World Problem', Journal of Personality and Social Psychology **15**, 101–118.

Kreps, D. and Wilson, R. (1982), 'Reputation and Imperfect Information', Journal of Economic Theory **27**(2), 253–279.

Krippendorff, K. (2004), Content Analysis: An Introduction to Its Methodology, Sage.

Krishnamurthy, S. (2002), 'The Multidemensionality Of Blog Conversations: The Virtual Enactment Of September 11', Internet Research **3**.

Kupiec, J., Pedersen, J. and Chen, F. (1995), A Trainable Document Summarizer, in 'SIGIR '95: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', ACM, Seattle, Washington, United States, pp. 68–73.

Kurland, O. and Lee, L. (2004), Corpus Structure, Language Models, and Ad-hoc Information Retrieval, in 'Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', ACM New York, NY, USA, Sheffield, UK, pp. 194–201.

Kurland, O. and Lee, L. (2005), PageRank Without Hyperlinks: Structural Re-ranking Using Links Induced by Language Models, in 'Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', ACM New York, NY, USA, Salvador, Brazil, pp. 306–313.

Lanagan, J. and Smeaton, A. (2007), SportsAnno: What Do You Think?, in 'RIAO'2007: Proceedings of the 8th Conference on Information Retrieval and its Applications', Pittsburgh, Pennsylvania, USA.

Lanagan, J. and Smeaton, A. F. (2009), Query Independent Measures of Annotation and Annotator Impact, in 'ESAIR '09: Proceedings of the WSDM '09 Workshop on Exploiting Semantic Annotations in Information Retrieval', ACM, Barcelona, Spain, pp. 35–38.

Langefors, B. (1973), Theoretical Analysis of Information Systems, Auerbach Publishers.

Larsen, B. and Ingwersen, P. (2002), The Boomerang Effect: Retrieving Scientific Documents Via The Network Of References and Citations, in 'SIGIR '02: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'.

Larsen, B. and Ingwersen, P. (2006), Using Citations For Ranking In Digital Libraries, in 'JCDL '06: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries'.

Lee, J. (1997), Analyses of Multiple Evidence Combination, in 'Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', Philadelphia, Pennsylvania, USA, pp. 267–276.

Lehmann, E. and D'abrera, H. (1975), Nonparametrics: Statistical Methods Based on Ranks, Holden-Day San Francisco.

Lerman, K. (2007), Dynamics Of Collaborative Document Rating Systems, in 'WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis', ACM, New York, NY, USA, pp. 46–55.

Li, R., Bao, S., Yu, Y., Fei, B. and Su, Z. (2007), Towards Effective Browsing of Large Scale Social Annotations, in 'WWW '07: Proceedings of the 16th International Conference on World Wide Web', ACM, New York, NY, USA, pp. 943–952.

Liu, H. and Zhang, H. (2005), 'A Sports Video Browsing and Retrieval System Based on Multimodal Analysis: Sportsbr', Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on **8**.

Liu, S., Zou, Q. and Chu, W. (2004), 'Configurable Indexing and Ranking For XML Information Retrieval', SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval .

Luhn, H. (1957), 'A Statistical Approach to Mechanized Encoding and Searching of Literary Information', IBM Journal of Research and Development **1**(4), 309–317.

Luhn, H. P. (1958), 'The Automatic Creation of Literature Abstracts', IBM Journal of Research and Development pp. 159–165.

MacRoberts, M. and MacRoberts, B. (1989), 'Problems of Citation Analysis', Scientometrics **36**(3), 435–444.

Manjunath, B., Salembier, P. and Sikora, T. (2002), Introduction to MPEG-7: Multimedia Content Description Interface, Wiley.

Maron, M. and Kuhns, J. (1960), 'On Relevance, Probabilistic Indexing and Information Retrieval', Journal of the ACM (JACM) **7**(3), 216–244.

Marsh, S. (1994), Formalising Trust as a Computational Concept, PhD thesis, Dept. of Computing Science and Mathematics.

Marshall, C. C. (1997), Annotation: From Paper Books To The Digital Library, in 'DL '97: Proceedings of the 2nd ACM International Conference on Digital Libraries', Philadelphia, Pennsylvania, United States, pp. 131–140.

Marshall, C. C. (1998), Toward an Ecology of Hypertext Annotation, in 'HYPERTEXT '98: Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia : Links, Objects, Time and Space—Structure in Hypermedia Systems', Pittsburgh, Pennsylvania, United States, pp. 40–49.

Marshall, C. C. and Brush, A. B. (2004), 'Exploring the relationship between personal and public annotations', Digital Libraries, Joint Conference on **7**(11), 349–357.

McCain, K. and Turner, K. (1989), 'Citation Context Analysis and Aging Patterns of Journal Articles In Molecular Genetics', Scientometrics **17**(1), 127–163.

McGill, M., Koll, M. and Noreault, T. (1979), An Evaluation of Factors Affecting Document Ranking by Information Retrieval Systems., Technical report, School of Information Studies, Syracuse University, New York.

Mealy, G. (1967), Another Look at Data., in 'Proceedings of the AFIPS Fall Joint Computer Conference', Vol. 31, Anaheim, California, USA, pp. 525–534.

Mika, P. (2007), 'Ontologies Are Us: A Unified Model of Social Networks and Semantics', Web Semantics: Science, Services and Agents on the World Wide Web **5**(1), 5–15.

Milgram, S. (1967), 'The Small World Problem', Psychology Today **2**(1), 60–67.

Millen, D. R. and Feinberg, J. (2006), Using Social Tagging to Improve Social Navigation, in 'Workshop on the Social Navigation and Community-Based Adaptation Technologies'.
URL: *http://www.sis.pitt.edu/%7Epaws/SNC_BAT06/crc/millen.pdf*

Motwani, R. and Raghavan, P. (1995), Randomized Algorithms, Cambridge University Press.

Nemrava, J., Buitelaar, P., Svátek, V. and Declerck, T. (2008), 'Text Mining Support for Semantic Indexing and Analysis of A/V Streams', OntoImage Workshop at LREC .

Newman, M. (2000), 'Models of the Small World', Journal of Statistical Physics **101**(3), 819–841.

Nielsen (2009), 'Twitter's Tweet Smell Of Success'.
URL: *http://blog.nielsen.com/nielsenwire/online_mobile/twitters-tweet-smell-of-success/*

Nov, O., Naaman, M. and Ye, C. (2008), What Drives Content Tagging: The Case of Photos on Flickr, in 'CHI '08: Proceeding of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems', ACM, New York, NY, USA, pp. 1097–1100.

NYTimes (1999), "RealNetworks Is Target of Suit In California Over Privacy Issue". The New York Times, Nov 9th, 1999.

O'Donovan, J. and Smyth, B. (2005), Trust in Recommender Systems, in 'Proceedings of the 10th International Conference on Intelligent User Interfaces', San Diego, California, USA, pp. 167–174.

O'Hara, K. and Sellen, A. (1997), A Comparison of Reading Paper and On-Line Documents, in 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', ACM New York, NY, USA, pp. 335–342.

O'Reilly, T. (2005), 'What is web 2.0'.
URL: *http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html*

O'Toole, C., Smeaton, A. F., Murphy, N. and Marlow., S. (1999), Evaluation of Automatic Shot Boundary Detection on a Large Video Test Suite, in 'CIR'99 - The Challenge of Image Retrieval: 2nd UK Conference on Image Retrieval'.

Ovsiannikov, I., Arbib, M. and McNeill, T. (1999), 'Annotation technology', International Journal of Human-Computer Studies **50**(4), 329–362.

Page, L., Brin, S., Motwani, R. and Winograd, T. (1998), 'The Pagerank Citation Ranking: Bringing Order To The Web'.

Paolillo, J. and Penumarthy, S. (2007), The Social Structure of Tagging Internet Video on del.icio.us, in 'IEEE International Conference on System Sciences', Vol. 40, IEEE, Hawaii, p. 1414.

Parker, M., Moleshe, V., De la Harpe, R. and Wills, G. B. (2006), An Evaluation of Information Quality Frameworks for the World Wide Web, in 'Proceedings of 8th Annual Conference on WWW Applications.', Bloemfontein, Free State Province, South Africa.

Pauly, D. and Stergiou, K. (2005), 'Equivalence of Results from Two Citation Analyses: Thomson ISI's Citation Index and Google's Scholar Service', Ethics in Science and Environmental Politics **2005**, 33–35.

Peritz, B. (1992), 'On The Objectives of Citation Analysis: Problems of Theory and Method', Journal of the American Society for Information Science **43**(6), 448–451.

Perneger, T. (1998), 'What's Wrong with Bonferroni Adjustments', British Medical Journal **316**(7139), 1236–1238.

Pool, I. and Kochen, M. (1978), 'Contacts and Influence', Social Networks **1**(1), 5–51.

Porter, M. (1980), 'An Algorithm for Suffix Stripping', Program: Electronic Library and Information Systems **40**(3), 211–218.

R Development Core Team (2004), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.

Redman, T. (1997), Data Quality for the Information Age, Artech House, Inc. Norwood, MA, USA.

Resnick, P., Kuwabara, K., Zeckhauser, R. and Friedman, E. (2000), 'Reputation Systems', Communications of the ACM **43**(12), 45–48.

Resnick, P. and Zeckhauser, R. (2002), 'Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System', The Economics of the Internet and E-Commerce **11**(2), 23–25.

Rieh, S. and Belkin, N. (1998), 'Understanding Judgment of Information Quality and Cognitive Authority in the WWW', Journal of the American Society for Information Science **35**, 279–289.

Ritchie, A., Robertson, S. and Teufel, S. (2008), Comparing Citation Contexts for Information Retrieval, in 'CIKM '08: Proceeding of the 17th ACM Conference on Information and Knowledge Management', ACM, Napa Valley, California, USA, pp. 213–222.

Robertson, S. (1977), 'The Probability Ranking Principle in IR', Journal of Documentation **33**(4), 294–304.

Robertson, S. E. and Sparck Jones, K. (1976), 'Relevance Weighting of Search Terms', Journal of the American Society of Information Science **27**, 129–146.

Robertson, S. E., Walker, S., Sparck Jones, K., Hancock-Beaulieu, M. M. and Gatford, M. (1994), 'Okapi at TREC-3', NIST Special Publication pp. 109–126.

Robertson, S. and Walker, S. (2000), 'Okapi/Keenbow at TREC-8', NIST Special Publication pp. 151–162.

Rosman, K. (1999), Booking Plugs on Amazon.com. Brill's Content, April 1999.

Rousseau, R. (2006), 'New Developments Related to The Hirsch Index', Science Focus **1**(4), 23–25.

Sabater, J. and Sierra, C. (2005), 'Review on Computational Trust and Reputation Models', Artificial Intelligence Review **24**(1), 33–60.

Sadlier, D. and O'Connor, N. (2005), 'Event Detection in Field Sports Video Using Audio-Visual Features and a Support Vector Machine', IEEE Transaction on Circuits and Systems for Video Technology **15**(10), 1225.

Salton, G. (1971a), 'Automatic Indexing Using Bibliographic Citations', Journal of Documentation **27**(2), 98–110.

Salton, G. (1971b), The SMART Retrieval System—Experiments in Automatic Document Processing, Prentice-Hall, Inc. Upper Saddle River, NJ, USA.

Salton, G. and Buckley, C. (1988), 'Term-Weighting Approaches in Automatic Text Retrieval', Information Processing and Management **24**(5), 513–523.

Salton, G. and Buckley, C. (1990), 'Improving Retrieval Performance by Relevance Feedback', Journal of the American Society for Information Science **41**(4), 288–297.

Salton, G. and McGill, M. (1986), Introduction to Modern Information Retrieval, McGraw-Hill, Inc. New York, NY, USA.

Salton, G., Wong, A. and Yang, C. (1975), 'A Vector Space Model for Automatic Indexing', Communications of the ACM **18**(11), 613–620.

Salton, G. and Yang, C. (1973), 'On the Specification of Term Values in Automatic Indexing', Journal of Documentation **29**(4).

Sanderson, M. (2008), 'Revisiting H Measured on UK LIS and IR Academics', Journal of the American Society for Information Science and Technology (JASIST) **59**(7), 1184–1190.

Sannomiya, T., Amagasa, T., Yoshikawa, M. and Uemura, S. (2000), 'A Framework for Sharing Personal Annotations on Web Resources Using XML', Information Technology for Virtual Enterprises, 2001. ITVE 2001. Proceedings. Workshop on pp. 40 – 48.

Schenkel, R., Crecelius, T., Kacimi, M., Michel, S., Neumann, T., Parreira, J. X. and Weikum, G. (2008), Efficient Top-K Querying Over Social-Tagging Networks, in 'SIGIR '08: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', ACM, New York, NY, USA, pp. 523–530.

Schilit, B. N., Golovchinsky, G. and Price, M. N. (1998), Beyond Paper: Supporting Active Reading With Free-Form Digital Ink Annotations, in 'CHI '98: Proceedings of the 16th SIGCHI Conference on Human Factors in Computing Systems in Computing Systems', Los Angeles, California, United States, pp. 249–256.

Schillo, M., Funk, P. and Rovatsos, M. (2000), 'Using Trust for Detecting Deceitful Agents in Artificial Societites', Applied Artificial Intelligence (Special Issue on Trust, Deception and Fraud in Agent Societies) **14**(8), 825–848.

Schreiber, M. (2007), 'Self-Citation Corrections for the Hirsch Index', Europhysics Letters **78**(3), 0295–5075.

Schreiber, M. (2008), 'The Influence of Self-Citation Corrections on Egghe's g Index', Scientometrics **76**(1), 187–200.

Scott, W. (1955), 'Reliability of Content Analysis: The Case of Nominal Scale Coding', Public Opinion Quarterly **19**(3), 321–325.

Seligman, A. (1997), The Problem of Trust, Princeton University Press.

Shardanand, U. and Maes, P. (1995), Social Information Filtering: Algorithms for Automating "Word of Mouth", in 'CHI '95: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', Denver, Colorado, United States, pp. 210–217.

Shipman, F., Price, M., Marshall, C., Golovchinsky, G. and Schilit, B. (2003), Identifying Useful Passages in Documents Based on Annotation Patterns, in 'Proceedings of the 7th European Conference on Digital Libraries', Vol. 2769, Trondheim, Norway, pp. 101–112.

Siegel, S. and Castellan, N. (1956), Non-Parametric Statistics for the Behavioural Sciences, McGraw-Hill New York.

Silverstein, C., Marais, H., Henzinger, M. and Moricz, M. (1999), 'Analysis of a Very Large Web Search Engine Query Log', **33**(1), 6–12.

Singhal, A., Buckley, C. and Mitra, M. (1996), Pivoted Document Length Normalization, in 'Proceedings of the 19th Annual International ACM Sigir Conference on Research and Development in Information Retrieval', Zurich, Switzerland, pp. 21–29.

Smeaton, A., Keogh, G., Gurrin, C., McDonald, K. and Sødring, T. (2003), 'Analysis of Papers From Twenty-Five Years of SIGIR Conferences: What Have We Been Doing For The Last Quarter of a Century?', ACM SIGIR Forum **37**(1).

Smeaton, A. and Quigley, I. (1996), Experiments on Using Semantic Distances Between Words in Image Caption Retrieval, in 'Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', Zurich, Switzerland, pp. 174–180.

Smyth, B., Balfe, E., Freyne, J., Briggs, P., Coyle, M. and Boydell, O. (2004), 'Exploiting Query Repetition and Regularity in an Adaptive Community-Based Web Search Engine', User Modeling and User-Adapted Interaction **14**(5), 383–423.

Sparck Jones, K. (1972), 'A Statistical Interpretation of Term Specificity and its Application in Retrieval', Journal of Documentation **28**(1), 1120.

Strong, D., Lee, Y. and Wang, R. (1997), 'Data Quality in Context', Communications of the ACM **40**(5), 103–110.

Tannenbaum, P. and Noah, J. (1959), 'Sportugese: A Study of Sports Page Communication', Journalism Quarterly **36**(2), 163–170.

Tatarinov, I., Viglas, S., Beyer, K., Shanmugasundaram, J., Shekita, E. and Zhang, C. (2002), 'Storing and Querying Ordered XML Using a Relational Database System', SIGMOD '02: Proceedings of the 2002 ACM SIGMOD international conference on Management of data .

Terveen, L., Hill, W., Amento, B., McDonald, D. and Creter, J. (1997), 'PHOAKS: A System For Sharing Recommendations', Communications of the ACM **40**(3), 59–62.

Tjaden, B. and Wasson, G. (1997), 'The Oracle of Bacon'.
URL: *http://oracleofbacon.org/*

Van House, N. (2002), Trust and Epistemic Communities in Biodiversity Data Sharing, in 'Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries', Portland, Oregon, USA, pp. 231–239.

Van Rijsbergen, C. (1979), Information Retrieval, Butterworth-Heinemann Newton, MA, USA, p. 16.

Vogt, C. C. and Cottrell, G. W. (1999), 'Fusion Via a Linear Combination of Scores', Information Retrieval **1**(3), 151–173.

Voorhees, E. (2000), 'Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness', Information Processing and Management **36**(5), 697–716.

Wand, Y. and Wang, R. (1996), 'Anchoring Data Quality Dimensions in Ontological Foundations', Communications of the ACM **39**(11), 86–95.

Wang, R., Lee, Y., Pipino, L. and Strong, D. (1998), 'Manage Your Information as a Product', Sloan management review **39**(4), 95.

Wang, R. and Strong, D. (1996), 'Beyond accuracy: what data quality means to data consumers', Journal of Management Information Systems **12**(4), 5–33.

Wang, R., Ziad, M. and Lee, Y. (2001), Data quality, Kluwer Academic Publishers Norwell, MA, USA, p. 2.

Wann, D., Metcalf, L., Adcock, M., Choi, C., Dallas, M. and Slaton, E. (1997), 'Language of Sport Fans: Sportugese Revisited.', Perception Motor Skills **85**(3 Pt 1), 1107–10.

Wasserman, S. and Faust, K. (1994), Social Network Analysis: Methods and Applications, Cambridge University Press.

Watts, D. and Strogatz, S. (1998), 'Collective Dynamics of 'Small-World' Networks', Nature(London) **393**(6684), 440–442.

Wiles, A. (1995), 'Modular Elliptic Curves and Fermat's Last Theorem', Annals of Mathematics **141**(3), 443–551.

Wilson, P. (1983), Second-Hand Knowledge: An Inquiry Into Cognitive Authority, Greenwood Press.

Windley, P. J., Daley, D., Cutler, B. and Tew, K. (2007), Using Reputation to Augment Explicit Authorization, in 'DIM '07: Proceedings of the 2007 ACM Workshop on Digital Identity Management', Fairfax, Virginia, USA, pp. 72–81.

Witten, I., Moffat, A. and Bell, T. (1999a), Managing Gigabytes: Compressing and Indexing Documents and Images, 2 edn, Academic Press/Morgan Kaufmann, chapter 3, p. 109.

Witten, I., Moffat, A. and Bell, T. (1999b), Managing Gigabytes: Compressing and Indexing Documents and Images, 2 edn, Academic Press/Morgan Kaufmann, chapter 3, pp. 113–115.

Wolfe, J. L. (2000), Effects of Annotations on Student Readers and Writers, in 'DL '00: Proceedings of the 5th ACM Conference on Digital Libraries', San Antonio, Texas, USA, pp. 19–26.

Wolff, J., Florke, H. and Cremers, A. (2000), 'Searching and Browsing Collections of Structural Information', Advances in Digital Libraries, 2000. ADL 2000. Proceedings. IEEE pp. 141 – 150.

Wu, X., Zhang, L. and Yu, Y. (2006), Exploring Social Annotations for the Semantic Web, in 'WWW '06: Proceedings of the 15th International Conference on World Wide Web', ACM, New York, NY, USA, pp. 417–426.

Xi, W., Lind, J. and Brill, E. (2004), Learning Effective Ranking Functions For Newsgroup Search, in 'Proceedings of the 27th Annual International Conference on Research and Developement in Information Retrieval', ACM Press New York, NY, USA, pp. 394–401.

Yang, K. Meho, L. (2006), 'Citation Analysis: A Comparison of Google Scholar, Scopus, and Web of Science', The American Society for Information Science and Technology **43**(1), 185.

Yu, B. and Singh, M. (2003), Searching Social Networks, in 'Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems', Melbourne, Australia, pp. 65–72.

Zacharia, G., Moukas, A. and Maes, P. (2000), 'Collaborative Reputation Mechanisms for Electronic Marketplaces', Decision Support Systems **29**(4), 371–388.

Zhai, C. (2001), 'Notes on the Lemur TFIDF Model'.

Zhang, J. and Ackerman, M. (2005), Searching for Expertise in Social Networks: A Simulation of Potential Strategies, in 'Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work', Sanibel Island, Florida, USA, pp. 71–80.

Zheng, Q., Booth, K. and McGrenere, J. (2006), Co-Authoring With Structured Annotations, in 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', Montréal, Québec, Canada, pp. 131–140.

Zhu, X. and Gauch, S. (2000), Incorporating Quality Metrics in Centralized/Distributed Information Retrieval On The World Wide Web, in 'SIGIR '00: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', Athens, Greece, pp. 288–295.

Ziman, J. (1968), 'Public Knowledge: The Social Dimension of Science', Cambridge: CUP .

Zimmermann, P. (1994), 'PGP User's Guide, Volume I: Essential Topics', Available on the WWW via ftp://ftp.pgpi.org/pub/pgp/2.x/doc/pgpdoc1.txt .

Zipf, G. (1949), Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology, Addison-Wesley Press.

Index