Identifying Useful and Important Information within Retrieved Documents

Piyush Arora Gareth J. F. Jones
ADAPT Centre, School of Computing,
Dublin City University, Dublin 9, Ireland
{parora,gfjones}@computing.dcu.ie

ABSTRACT

We describe an initial study into the identification of important and useful information units within documents retrieved by an information retrieval system in response to a user query created in response to an underlying information need. This study is part of a larger investigation of the exploitation of useful and important units from retrieved documents to generate rich document surrogates to improve user search experience. We report three user studies using a crowdsourcing platform, where participants were first asked to read an *information need* and *contents of a relevant document* and then to perform actions depending on the type of study: i) write important information units *(WIIU)*, ii) highlight important information units *(HIIU)* and iii) assess importance of already highlighted information units *(AIHIU)*. Further, we discuss a novel mechanism for measuring similarities between content annotations. We find majority agreement of about 0.489 and pairwise agreement of 0.340 among users annotation in the AIHIU study, and average cosine similarity of 0.50 and 0.57 between participant annotations and documents in the WIIU and HIIU studies respectively.

1. INTRODUCTION

Document surrogates are a primary way in which users interact with potentially interesting documents in information retrieval (IR) applications. Classically document surrogates are intended to enable users to assess the potential relevance of a document, rather than to provide important and useful information from the document itself to directly address the information need underlying the current search or more generally to improve the user's topical knowledge. The primary focus in IR has been on optimizing topical *relevance* by retrieving documents deemed *relevant* to the user's information need [5]. More recently it has been proposed that research and evaluation in IR should take a more sophisticated view of the objective of IR to include measuring features such as *utility* of retrieved information, the user's knowledge, or some combination of these parameters [1, 12]. In response to this IR researchers have begun to look beyond traditional topical relevance to evaluation metrics such as *usefulness*, *effort* and *readability* [3, 10, 16] to capture and satisfy user needs effectively but at the *document* or *IR system* level. However, in general only parts of a document will be of interest to the user. Recognising this there has been recent work to examine this topic. For example, Habernal et al. focused on finding sentence level relevant information within a document. In this work the authors specified detailed guidelines of relevance for each topic separately [4], which makes the task of sentence level annotation quite complex and laborious in general. Also of interest to this topic is previous work on XML and passage retrieval [6, 14] which looked at passages and the sentence level, but focused only on topical relevance. To the best of our knowledge no previous work has explored sentence level or other sub-document units to measure metrics such as usefulness and importance of text within a document.

We believe identifying useful and important information units at the sub-document level can be used in the generation of richer surrogate representations of documents to facilitate more meaningful user engagement with retrieved information in search [2, 15]. These useful and important information units offer alternative mechanisms to provide answers to user information needs, as in the case of question-answering systems and presenting information cards to the users [13]. As part of our investigation of this proposal, in this paper we study annotation of important and useful information within documents judged relevant to an information need. In particular we focus on consensus and agreement between users' annotations and judgments of annotated content.

2. EXPERIMENTAL INVESTIGATION

In this section, we introduce the design of the user studies and dataset used for our experiments.

2.1 Users study design

Participants were presented with a series of user information needs and a single relevant document for each one. Participants were recruited through the *Prolific* crowdsourcing platform[1], where each participant worked with just one of the interaction studies: WIIU, HIIU or AIHIU. Our study is composed of three types of user studies as follows:

Study-1 (WIIU): Re-writing important information units: Participants identify and re-write in their own words important and useful information units from within the document that satisfies the information need. We adopt the definition of information unit (iUnit) from the NTCIR Mobile click task [8], where information units are defined as relevant and atomic pieces of information, where *Relevant* means that a textual unit provides useful factual information to the user; *Atomic* means that a textual unit cannot be broken down into multiple units without loss of the original semantics.

CHIIR '17, March 07-11, 2017, Oslo, Norway
© 2017 ACM. ISBN 978-1-4503-4677-1/17/03...$15.00
DOI: http://dx.doi.org/10.1145/3020165.3022154

[1] https://www.prolific.ac/

Study-2 (HIIU): Highlighting important information units: Following previous work on XML and Passage retrieval [6, 14], participants highlight important and useful information units from within the document that satisfies the information need.

Study-3 (AIHIU): Assessing already highlighted information units: Participants assess already highlighted information units on a scale of [1-4]. The first author of this paper manually identified and highlighted topically related textual units from the documents to be categorized by the users between 4 classes of relevance and importance: i) C1: Highly relevant and important, ii) C2: Fairly relevant and important, iii) C3: Slightly relevant and important and iv) C4: Neither relevant nor important.

Research Contribution: The main contribution of this paper is the study of three different methods of how people analyze information beyond the document level to find and assess important and useful units within a document. We find similarity values between user annotations when they highlight and write information units for a given relevant document and agreement values among users when assessing information units. We also discuss a novel technique for measuring similarities between human annotations and retrieved documents.

Our studies focus on the following specific research questions:

RQ-1: How can we compare and measure user annotations? (the WIIU and HIIU studies)

RQ-2: What is the agreement among users while assessing information units already marked in a document? (the AIHIU study)

2.2 Dataset

We used data from the TREC 2012 session track for our study [7]. We selected 3 information needs for this dataset and at random one relevant document from the *qrels* for each of the three information needs. Since this is a comprehensive and cognitively intensive task for our participants, we opted to concentrate on detailed descriptive analysis of a small number of documents for this initial study, with the main goal of analyzing important and useful richer units within a document.

Differences in the user's topic familiarity can influence their search behaviour [9], thus to ensure people are familiar with the topics, we carefully chose following three simple and generic topics from the TREC data set:

– T0: Wedding Traditions

– T1: Smoking Cessation

– T2: Junk Food

2.3 Study Procedures

Participant Training: To carry out the user studies, topics were organised and always presented in the same order as described below for user training and test data collection:

WIIU: One sample topic (T0) was used to familiarize participants with the task and the other two topics to collect test annotations [2] (T1 and T2).

HIIU: One sample topic (T0) was used to familiarize the participants with the task and the other two topics to perform annotations[2] (T1 and T2).

AIHIU: All three topics were used to perform annotations[2] (T0, T1 and T2).

Data Collection: Annotations[2] were collected using Prolific crowdsourcing platform. Table 1 shows the demographics of the participants.

Study type	Users	Age Range	Demographics	Nationality
WIIU study	7	21-48	5 M & 2 F	4 US & 3 UK
HIHU study	7	21-35	3 M & 4 F	7 UK
AIHIU study	7	20-43	4 M & 3 F	5 UK & 2 US

Table 1: Participants Demographics Information

All participants were native English speakers. In accordance with standard crowdsourcing practice for this type of work, they were paid between 8-9 euros on a per hour basis.

After conducting a pilot run with 5 volunteers locally within our lab, we carried the user studies in two phases:

Phase-1 *WIIU and HIIU studies*: For each study we recruited seven multiple annotators as described in Table 1, to read documents and find useful and important information at textual level. Each study had 3 information needs, one information need was used for training to get familiar with the task and the other two were used to obtain the annotations as discussed in Section 2.2.

Phase-2 *AIHIU study*: In this method we asked the annotators to assess already highlighted textual units as discussed in Section 2.1.

3. EXPERIMENTS AND RESULTS

In this section we summarize the data collected and the results obtained in our three user studies.

3.1 Data Collection & Analysis

We collected following information from users:

i) Information units identified by the user (**WIIU and HIIU studies**)

ii) Assessment of highlighted information units on a scale of [1-4] with reasons (**AIHIU study**).

We carried out data analysis for the information units collected for the WIIU and HIIU studies. While writing information units in the WIIU study, participants rephrased the textual information, and in the HIIU study they freely highlighted the textual content. Thus calculating normal agreement across users is hard since the text boundaries are flexible and quite variable in nature. Hence we calculated word overlap and similarity between the information units and the document to analyze users response. This gives us a rough approximation of the consensus between participant annotations. We explore two different mechanisms to analyze information units for a given document:

a) *Quantitative approach:* We hypothesize that calculating the overlap of words, tokens and common nouns between participant annotations and documents can give a meaningful indication of agreement across participant responses. Thus for each participant we calculate overlap between the combined information units and the document in terms of number of words (W_O), tokens (U_W) and Nouns (N_N). We used NLTK toolkit [3] to perform part of speech tagging to extract Nouns from the document and information units.

b) *Qualitative approach:* We calculated the following two similarity scores between the information units and document to analyze user's annotations effectively.

1) *Cosine similarity (COS_S):* The cosine similarity between the document and the information units combined together.

2) *Word2Vec similarity (W2V_S):* We used *Word2Vec* vectors,

[2] "annotation" is used interchangeably, depending on the user study it means one of the three: find and write information units, highlight information units or assess already highlighted information units

[3] http://www.nltk.org/

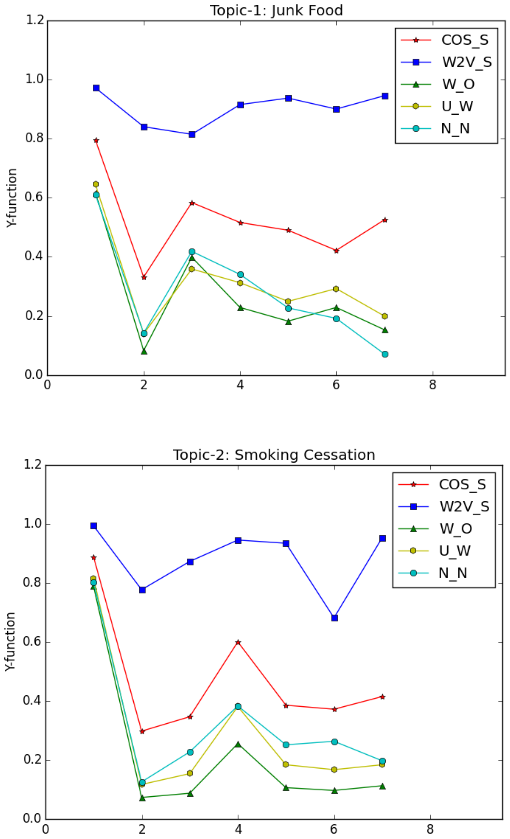

Figure 1: WIIU study

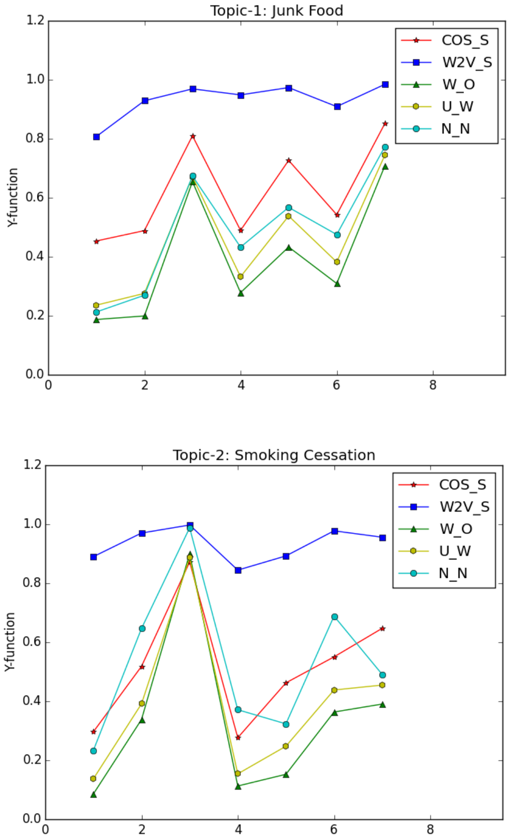

Figure 2: HIIU study

distributional representations of words, computed using neural networks [11] to calculate the cosine similarity between the centroid of all the word vectors in the document and the centroid of all the word vectors in the information units combined together. Word vectors were obtained using the Gensim toolkit [4].

3.2 Results

This section details the results of our three user studies. Figure 1 and 2 represent the participant annotation similarity and overlap scores for the WIIU and HIIU studies, across 7 users for T1 and T2 respectively. Table 2 details average users annotation similarity and overlap scores for WIIU and HIIU studies. Using different factors such as COS_S, W2V_S, W_O, U_W and N_N for measuring the similarities provide a holistic method to evaluate the annotations as looking at just one factor e.g. COS_S can be biased towards the length and number of information units.

Table 3 shows the agreement values for the AIHIU study. For each annotation we had 7 annotators and 4 classes. Because of the variance in terms of the distribution of classes and multiple annotators the scores of Fleiss Kappa, Pairiwse Cohen's Kappa and Krippendoff's kappa were quite low 0.012, 0.054 0.015 respectively. Thus we show only the majority agreement values in Table 3. These indicate the scores when majority of the users (>=4) agreed with the annotation class. We also tried reducing the 4 class relevance scale to 3 classes by combining class C2 and C3 i.e. "fairly relevant and important", and "slightly relevant and important". Agreement scores between three classes are also shown in Table 3.

Study type	COS_S	W2V_S	W_O	U_W	N_N
WIIU_T1	0.523	0.903	0.270	0.313	0.285
WIIU_T2	0.472	0.879	0.217	0.286	0.321
HIIU_T1	0.623	0.931	0.395	0.454	0.486
HIIU_T2	0.517	0.932	0.334	0.387	0.534

Table 2: Comparison across different methods

4. ANALYSIS AND DISCUSSION

In this section we analyze the results of our user studies.

4.1 Study based Observations

WIIU study: More text written as information units results in better cosine similarity due to the increase in overlap of words. However, word2vec similarity results indicate that even with a lower number of units the similarity is higher than 80%. Some participants also rephrase the units, thus the cosine similarity and overlap scores decrease but the word2vec similarity is still high which is evident in Figure 1.

HIIU study: In the easier task of highlighting participants actually highlighted most of the text. The results here have quite a lot of noise due to the flexibility of free choice of the starting and ending points of the highlighted text with no constraints on the length of the textual piece to be highlighted. Thus it may be more apt to get annotators to mark documents at fixed sentence level and be more specific in the annotation guidelines in the future.

Results are more consistent between users when they write information units as compared to when they highlight information units, this is indicated by the frequent ups and down in the lines plot as shown in Figure 1 and 2. In the WIIU and HIIU studies partici-

[4] https://radimrehurek.com/gensim/models/word2vec.html

Study type	No. of units	4 class relevance	3 class relevance
AIHIU_T0 (maj agr)	14	0.500	0.787
AIHIU_T1 (maj agr)	19	0.421	0.840
AIHIU_T2 (maj agr)	14	0.571	0.857
Average (maj agr)	47	0.489	0.829
Pair Agr	47	0.340	0.430

Table 3: Agreement for the AIHIU studies, Maj Agr: means majority agreement, Pair Agr: means pairwise agreement, calculating agreement of 2 annotators at a time and averaging over all the combinations.

pants reported that it is quite challenging to identify important and useful text when they read documents. Some people encounter new information and find everything useful and important, whereas others who know about the topic can be too hard and restrictive when judging the importance and usefulness of textual information without strict guidelines.

AIHIU study: The majority agreement averaged across three topics is 0.489 and the pairwise agreement is 0.340 as shown in Table 3. This is on similar lines to the agreement values reported for XML retrieval [6, 14]. Assessing information units as important and useful introduces a subjective aspect depending on the participant's prior knowledge, further assessment of these units at 4 levels of categories makes it challenging to get satisfactorily agreement between multiple annotators.

4.2 Issues with the study

We are aware of three shortcomings with the study:

1) The number of users was limited, we had 7 data points for each task. On average participants took about 15-20 minutes to complete each task. This limited us to have each participant working with just one of the interaction studies: WIIU, HIIU or AIHIU.

2) This is a preliminary study with 3 topics and 1 relevant document for each of the three topics. Further studies needs to be carried out with more topics and documents to draw effective conclusions.

3) Flexibility of choosing starting and ending points while highlighting text makes it challenging to calculate agreement between multiple annotators and needs further exploration.

5. CONCLUSION AND FUTURE WORK

In this work, we investigated the location and analysis of important and useful information within relevant documents using three different user study methods. We investigated two research questions:

RQ-1: How can we compare and measure user annotations (WIIU and HIIU studies)? – In this work, we calculated a combination of different measures such as textual overlap (including words, tokens and noun), cosine and word2vec similarity to compare and measure user's annotations, as just using the cosine similarity can be biased towards the length and number of information units. These measures collectively can give an approximation of the overlap between user annotations where it is difficult to measure direct agreement between annotations.

RQ-2: What is the agreement among users while assessing information units already marked in a document (AIHIU study)? – We calculated the majority agreement which is about 0.489 and the pairwise agreement which is about 0.340 among the users while categorizing information units into one of the four classes.

The objective of this study was to explore the potential to analyze information within documents beyond relevance to measures like importance and usefulness of information within documents. Though the number of data points is limited in this initial study, they indicate that it is realistic to annotate beyond document level to evaluate metrics such as useful, important and relevant information. The results and analysis indicate that we can look beyond the document level, but that it would be more practical to work with fixed boundary units such as sentence level rather than free annotation of textual units. In future work, we will explore the topics opened up in this study further with larger numbers of participants. Additionally, the results of this work will contribute to our broader objective of creation of richer document surrogate and summaries, and effective presentation of information to users to promote for effective search and engagement, and emerging areas such as improving learning through search.

Acknowledgment: This research is supported by Science Foundation Ireland (SFI) as a part of the ADAPT Centre at Dublin City University (Grant No: 12/CE/I2267).

6. REFERENCES

[1] A. Al-Maskari and M. Sanderson. A review of factors influencing user satisfaction in information retrieval. *JASIST, 2010*, 61(5):859–868.

[2] P. Arora and G. J. F. Jones. Position paper: Promoting user engagement and learning in search tasks by effective document representation. In *Proceedings of SAL workshop, SIGIR 2016*.

[3] M. Cole, J. Liu, N. Belkin, R. Bierig, J. Gwizdka, C. Liu, J. Zhang, and X. Zhang. Usefulness as the criterion for evaluation of interactive information retrieval. *Proc. HCIR*, pages 1–4, 2009.

[4] I. Habernal, M. Sukhareva, F. Raiber, A. Shtok, O. Kurland, H. Ronen, J. Bar-Ilan, and I. Gurevych. New collection announcement: Focused retrieval over the web. In *Proceedings of SIGIR 2016*, pages 701–704.

[5] K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of SIGIR 2000*, pages 41–48.

[6] J. Kamps, J. Pehcevski, G. Kazai, M. Lalmas, and S. Robertson. Inex 2007 evaluation measures. In *International Workshop of the Initiative for the Evaluation of XML Retrieval*, pages 24–33. Springer, 2007.

[7] E. Kanoulas, B. Carterette, M. Hall, P. Clough, and M. Sanderson. Overview of the TREC 2012 Session Track. 2012.

[8] M. P. Kato, M. Ekstrand-Abueg, V. Pavlu, T. Sakai, T. Yamamoto, and M. Iwata. Overview of the ntcir-11 mobileclick task. In *NTCIR*, 2014.

[9] D. Kelly and C. Cool. The effects of topic familiarity on information search behavior. JCDL '02, pages 74–75. ACM, 2002.

[10] J. Mao, Y. Liu, K. Zhou, J.-Y. Nie, J. Song, M. Zhang, S. Ma, J. Sun, and H. Luo. When does relevance mean usefulness and user satisfaction in web search? In *Proceedings of SIGIR 2016*, pages 463–472.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[12] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part iii: Behavior and effects of relevance. *JASIST, 2007*, 58(13):2126–2144.

[13] M. Shokouhi and Q. Guo. From queries to cards: Re-ranking proactive card recommendations based on reactive search history. In *Proceedings of SIGIR 2015*, pages 695–704.

[14] A. Trotman and S. Geva. Passage retrieval and other xml-retrieval tasks. In *Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology*, pages 43–50, 2006.

[15] R. W. White, J. M. Jose, and I. Ruthven. Using top-ranking sentences to facilitate effective information access. *JASIST, 2005*, 56(10):1113–1125.

[16] E. Yilmaz, M. Verma, N. Craswell, F. Radlinski, and P. Bailey. Relevance and effort: An analysis of document utility. In *Proceedings of CIKM 2014*, pages 91–100.