SIGIR: Scholar vs. Scholars' Interpretation

James Lanagan and Alan F. Smeaton
CLARITY: Centre for Sensor Web Technologies
Dublin City University
Dublin, Ireland
{jlanagan, asmeaton}@computing.dcu.ie

ABSTRACT

Google Scholar allows researchers to search through a free and extensive source of information on scientific publications. In this paper we show that within the limited context of SIGIR proceedings, the rankings created by Google Scholar are both significantly different and very negatively correlated with those of domain experts.

Categories and Subject Descriptors

H.1.2 [**MODELS AND PRINCIPLES**]: User/Machine Systems—*Human information processing*

General Terms

Algorithms, Experimentation, Human Factors

1. INTRODUCTION

The launch of Google Scholar[1] (GS) in late 2004 meant that scholars were suddenly provided with a free and extensive source of scientific information for searching and citing. Even though this resource is free, it has been shown to compare well with the performance of paid indexes such as the Web of Science[2] [1]. But one of the largest criticisms levelled against GS is its lack of transparency in its ranking methods: *"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."*

In this paper we use 10 years of SIGIR conference papers as a testbed for comparing expert rankings vs. GS rankings, and we reach some surprising conclusions.

2. IMPLEMENTATION

The first study of SIGIR conference proceedings was performed as part of the 25th anniversary celebrations of the SIGIR conference [4], and later extended for the 30th year of the conference [2]. These studies focused on the closed collection of SIGIR papers and have not, to our knowledge, taken into account any citations of SIGIR papers by papers either external or internal to SIGIR. That previous work [4] also clustered the first 25 years of SIGIR proceedings into several distinct and reoccurring topics. We have now identified a new larger set of topics which cover the years (1997-2007). Using the session names from within each SIGIR conferences together with the cluster names from [4], we identified 12 topics (Table 1) covering long-standing interests of the IR community, as well as new interests such as spam, and these are used in our experiments.

Table 1: Number of documents ranked by our experts for each of the topics chosen.

Topic	*Query*	*Documents Ranked*
Collaborative Filtering (CF)	"collaborative filtering"	10
Distributed IR (DR)	"distributed retrieval"	8
Document Clustering (DC)	"document clustering"	10
Image Retrieval (IR)	"image retrieval"	11
Language Modeling (LM)	"language model"	12
Latent Semantic Indexing/Analysis (LS)	"latent semantic"	12
Linkage Analysis (LA)	"link analysis"	10
Question Answering (QA)	"question answer"	9
Relevance Feedback (RF)	"relevance feedback"	10
Spam (S)	"spam"	6
Text Summarisation (TS)	"text summarization"	9
Topic Distillation (TD)	"topic distillation"	8

To create a dataset for searching we built an extended SIGIR citation graph from a 10 year window (1997-2007) of full papers in SIGIR proceedings[3]. Within this there are over 4,000 authors, ∼770 SIGIR publications and an additional ∼2,100 non-SIGIR publications which cite these SIGIR articles. We have calculated PageRank scores for every document, allowing us to generate ranked lists.

We used our 12 topics to generate a list of documents to present to experts by combining the top 30 documents returned from a query against GS (a restricted query returning papers from the SIGIR proceedings published between 1997-2007) with a ranked list returned for the same query against our extended SIGIR citation network, and then using the top-ranked papers that appeared in both lists.

We then asked 14 expert users from 3 different university information retrieval research groups[4] to provide rankings for each topic's list of documents. For each topic, experts were given the first page *only* from each paper and asked to provide a ranking suitable for a novice research student

[1]http://scholar.google.com/intl/en/scholar/about.html
[2]http://isiknowledge.com/

[3]Our time-window was defined as a result of the limited availability of machine-readable documents prior to 1997.
[4]Dublin City University, University College Dublin, and Glasgow University.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

interested in the topic. Broad queries against Google Scholar had specifically been used to simulate a novice user searching for relevant papers about a topic. An explanation of the rankings was requested, as well as a topic expertise rating of 1 (*"I have had no real experience of this topic"*) to 5 (*"I am knowledgeable in this topic"*). The main reasons given for ranking papers highly were author, institution, scope, content, and year of publication[5].

While the reasons that experts gave for highly ranked documents overlapped greatly, the rankings themselves were not uniform. Overall we collected 1,082 document judgements with an average of 7 judgements per paper, and 6 of the 14 experts ranked all 12 topics.

We used the Kendall coefficient of concordance (W) to measure inter-rater agreement [3], showing significant agreement within each topic's expert rankings[6]. This enables us to use the median expert rank of each paper within a topic to create a new combined ranking for that topic. In cases where the median of two papers' ranks are equal, the mean ranks are used to decide the ordering.

3. COMPARING SCHOLAR TO EXPERTS

Correlations between the combined experts' rankings and those created by GS lead us to believe that the rankings that GS is modelling are far from expert: GS's algorithm seems to use features available through direct analysis of the papers, quite the opposite to expert assessors who may call upon past experience and prior knowledge — prior knowledge that increases with the level of self-determined expertise of the ranking expert.

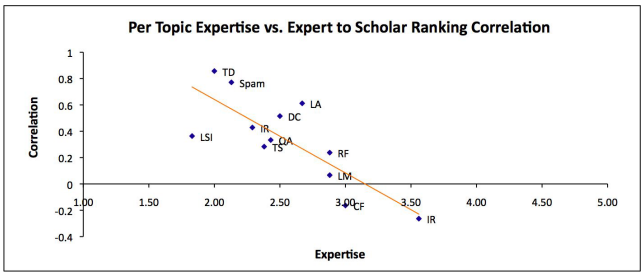

Figure 1: The per-topic correlation of expert and scholar rankings.

This can be seen in the decrease in correlation between per-topic expert rankings and GS rankings as the average (mean) expertise of that topic's experts increases (Figure 1). The reason for issuing broad and non-specific queries to GS as shown in Table 1 is to simulate the inexperienced user who comes to our experts with a selection of papers and no clear idea of their relative values. We expected the GS vs expert rankings to correlate well with each other regardless of expertise. Instead, there is a -0.7922 correlation between 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 their ranking.*

[5]No expert was asked to rank papers that they had authored, nor any from their own institution.

[6]This measure was used due to the ordinal nature of our data. Although the experts created the rankings independent of each other, the ranking a document receives is not independent of the other documents.

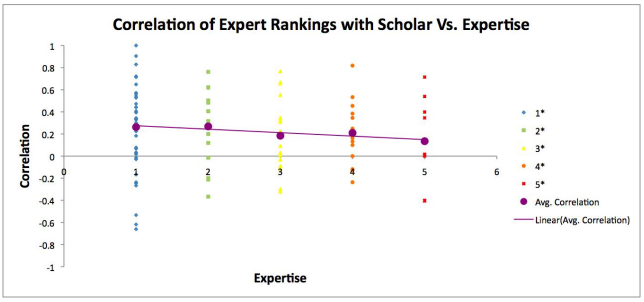

Figure 2: The correlation of per-expert and scholar rankings, divided into differing levels of expertise.

If we now look at the per-expert correlations with the GS ranking as shown in Figure 2, we see that whilst the correlation is not as strongly negative as on a per-topic basis it remains negatively correlated. The graph does not use within-topic agreements of rankings amongst the experts, looking only at the level of agreement between each self-assigned expertise level's ranking and that of GS. It is interesting nonetheless that the divergence of expertise and GS is repeated at this level also.

4. CONCLUSIONS

While it may be argued that the ranking Google Scholar provides is designed to best fit user expectation and need, we do not feel that this ranking is optimal for broad topics. It appears from our results that the expectation being met is that of someone unfamiliar with the area being queried, and this is less desirable than having an expert rank output. This is an interesting observation given the continued rise of Google Scholar as a source for researchers, and one we feel is worthy of further investigation.

Acknowledgments

This work is supported by Science Foundation Ireland under grant number 07/CE/I1147.

5. REFERENCES

[1] A. Harzing and R. van der Wal. Google Scholar: The Democratization of Citation Analysis? *Ethics in Science and Environmental Politics*, 8(1):61–73, Jan 2007.

[2] D. Hiemstra, C. Hauff, F. Jong, and W. Kraaij. SIGIR's 30th Anniversary: An Analysis of Trends in IR Research and The Topology of Its Community. *ACM SIGIR Forum*, 41(2), Dec 2007.

[3] M. Kendall and B. Smith. The Problem of m Rankings. *Annals of Mathematical Statistics*, 10(3):275–287, 1939.

[4] A. Smeaton, G. Keogh, C. Gurrin, K. McDonald, and T. Sødring. 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), Apr 2003.