SIGIR's 30th anniversary: an analysis of trends in IR research and the topology of its community

Djoerd Hiemstra[1], Claudia Hauff[1], Franciska de Jong[1,2], and Wessel Kraaij[3]
[1]University of Twente [2]Erasmus University [3]TNO-ICT
{*d.hiemstra, c.hauff, f.m.g.dejong*}*@utwente.nl* *wessel.kraaij@tno.nl*

Abstract

This paper presents an analysis of all SIGIR proceedings to date in order to summarize what IR researchers discussed over the years, where they are from, and whether subcommunities can be identified, determined by co-authorship.

1 Introduction

In a period where social communities have become an object of active IR research, we set out to analyze the SIGIR community itself. The first investigation of this kind was performed in 2002 by Alan Smeaton et al. [5] in the context of the 25th anniversary of ACM SIGIR conference, which concentrated on a co-authorship analysis. A similar study was done by Mario Nascimento et al. [4] for ACM SIGMOD. Smeaton's work was inspired by the folkore of computing Erdös numbers in mathematics. These numbers are defined as the shortest path to Paul Erdös in a graph based on pairwise co-authorship relations. A person's Erdös number is 1 if he or she has published a paper with Erdös. It is 2 if he or she has published with someone who has published with Erdös, and so on. Although Paul Erdös was a well respected mathematician, his fame was mainly due to his prolific scientific output and his numerous co-authors [5] making him an important person in the collaboration graph of the mathematics community. A more objective way to determine the most influential person in a collaboration graph is to compute the shortest average path length to other persons in the graph. This methodology is the driving force behind the 'Oracle of Bacon', which analyzes the cast lists of movies in the internet movie database. Since the database is updated with new movies regularly, the centre of the hollywood universe changes over the years. At the time when Smeaton et al. performed their analysis, Kevin Bacon was already superseded by Christopher Lee as Hollywood's most central actor. In 2002, the centre of the SIGIR universe was Christopher Buckley and he was honoured with the "Christoper Lee Award". In 2007, the centre of the Hollywood universe is "Rod Steiger". Given the fact that the centre of a collaboration graph is time dependent, the collaboration graph is incrementally updated, we considered it appropriate to compute the current (2007) centre of the SIGIR universe and the winner of the award, Wensi Xi, was announced during the SIGIR 2007 banquet.

As part of SIGIR's 30th anniversary celebrations, a web application was launched, enabling members of the SIGIR community to compute their "Xi number": `http://www.sigir2007.org/search`. In addition users can enter their favorite IR topic to search in the titles and abstracts of 30 years of SIGIR proceedings and find authors (expert search), periods and geographical locations associated with their search term. All experiments were carried out using the PF/Tijah system for XML search [2]. In this paper we will describe the co-authorship analysis and report on the experiments carried out to capture the trends in IR topics, and in the geographical background of the papers. In Section 2 we analyse what topics have been discussed in the SIGIR proceedings over time. Section 3 analyses in what countries the authors were based. Section 4 contains the co-author analysis, and finally, Section 5 concludes the paper.

2 *What* have SIGIR authors been writing about, and *when*?

To find out what the IR community has been writing about, we built a simple language model for each year in our data, following the temporal language models approach suggested by De Jong et al. [1] (model B). For each particular year, the approach defines a standard language model. Given a particular word, however, we now have probabilities of that word for each single year. We call these probabilities the temporal word profile. Figure 1 shows the temporal word profile (using raw frequencies instead of probabilities) for the word *TREC*.

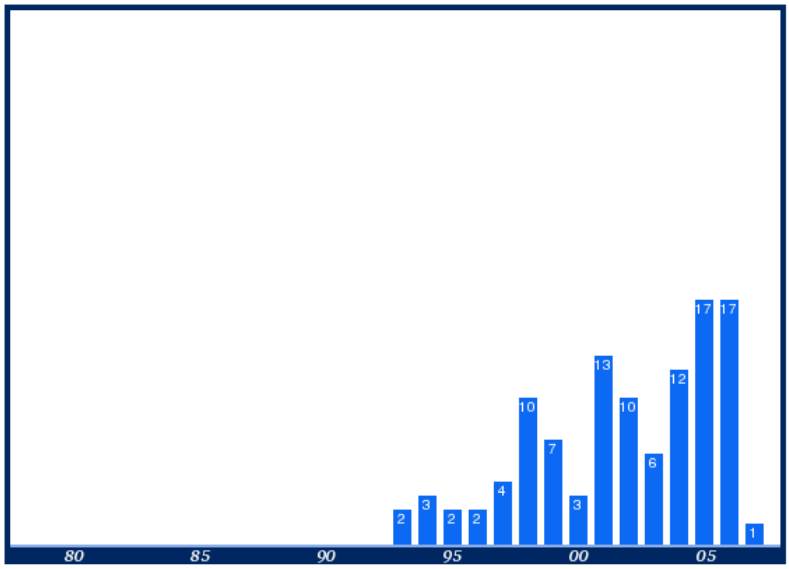

Figure 1: Temporal profile for the query *TREC*

Obviously, the number of publications that mention TREC as term has been increasing over the years. We might therefore conclude from this graph that TREC is a trendy word, or a buzz word: It seems authors should mention TREC these days to get their paper accepted at SIGIR. To measure a word's "trendiness" we calculated the correlation between a term's probability in the language model, and the year of publication. Terms with a positive

correlation have been used more recently, but have not been used a lot in the past. Terms with a negative correlation have been used in the past, but have not been used recently: words of nostalgia. We used the standard Pearson's correlation coefficient. One might argue that its underlying assumption of normality is not fully adequate given the discrete nature of textual data, but the language model probabilities behave similarly to continuous measures and we observed that the data approximates the normal distribution quite well. We removed stop words and added phrases from a domain-specific list that contains for instance *question answering* as one phrase.

word	correlation
bibliographic	−0.65
computer	−0.56
data base	−0.47
environment	−0.47
program	−0.44
implementation	−0.43
records	−0.41
file	−0.41

word	correlation
classification	0.78
trec	0.77
text	0.75
web	0.74
significant	0.70
question answering	0.67
cross language	0.66
latent semantic	0.64

Table 1: a) Nostalgic words b) Trendy words

Table 1 contains the words that show a negative correlation with the publication year, and the words that show a positive correlation with the publication year. The table shows that in the past, we used to search for *bibliographic* information, using a *computer*, a *data base* and *records*. Some words, e.g., *computer*, *program* and *implementation*, do not seem to be nostalgic at all, but apparently, they were used a lot in the past, but not that much anymore today. Maybe computers, programs and implementations have become such obvious attributes of the research set-up described that mentioning them is felt as redundancy, which could explain why these terms do not occur anymore these days? Surprisingly, at least to us, *classification* is the most trendy word in 30 years of SIGIR, maybe because of the increase in the use of machine learning methods for ranking. The occurrence of words such as *trec* (you have to do a TREC experiment these days), *web* (there was no world wide web in 1978) and *significant* (you better show your experimental results are significantly different these days) are no surprise, as well as *question answering*, *cross language* and *latent semantic*. Interestingly, if we consider stopwords as well, then two of the most positively correlated words are *we* (0.77) and *our* (0.74). This illustrates that the style used in research papers changed over years as well: 30 years ago, researchers would not use the first person plural, as in *"We ran TREC topics 551 to 599..."*, but instead kept their writing more impersonal as in *"TREC topics 551 to 599 were run..."*.

3 *Where* do SIGIR authors come from?

Another parameter for which we performed an analysis is the geographical origin of the papers published. Contribution from some 41 countries have been published, but there are huge differences in the spread of productivity over the various countries. The absolute all time number one contributor is the USA. In terms of productivity this country is so much a category of its own that if we would have included their 595 papers, the table below would

had to be split over two SIGIR Forum pages in order not to blur the distinction between the other countries.

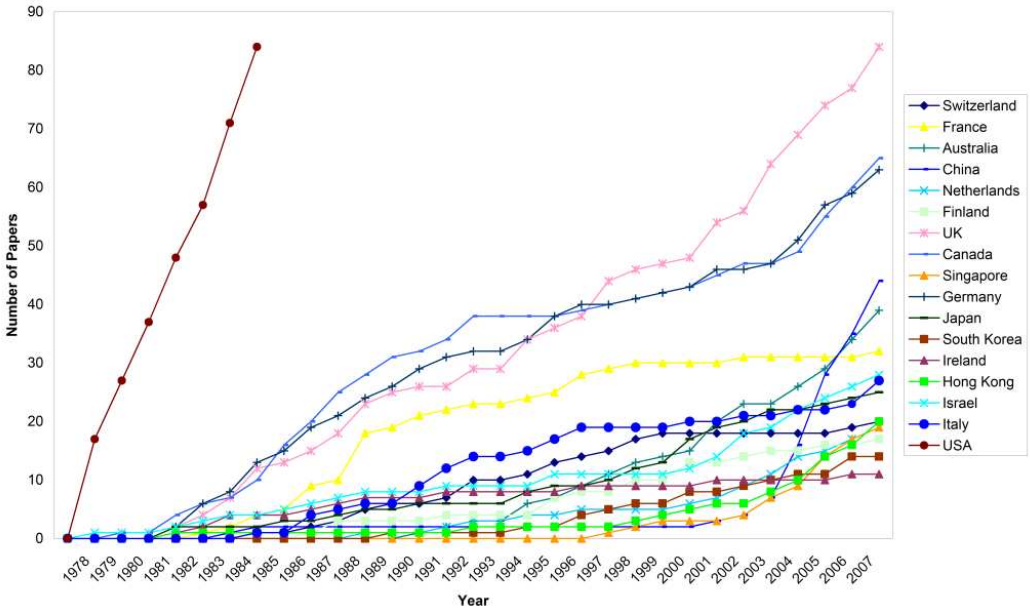

Figure 2: Countries with 10 or more papers

The table shows the figures for all other countries with ten or more papers in the period 1978–2007. One can distinguish three groups: the subtop consisting of the UK, Canada, and Germany; the subsubtop, a category which for many years consisted solely of France, but recently was entered by runner-ups China and Australia; and the rest. The latter category is where one can find a few other fast risers.

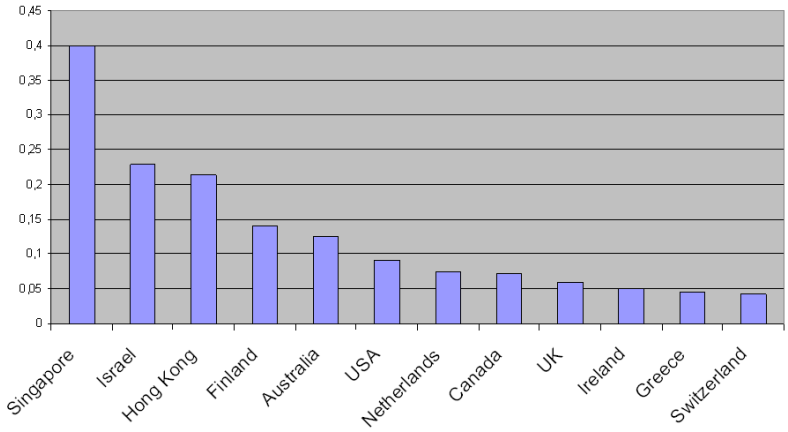

Figure 3: Number of papers per year per million inhabitants (1998–2007)

If we however normalise the number of papers produced by a country by its number of inhabitants, then the USA are no longer the biggest contributor to SIGIR. Figure 3 shows these numbers for the last 10 years. Interestingly, all countries in the top 12 have organised

SIGIR at some point, or will organise SIGIR in the near future, except for Hong Kong and Israel. Maybe –without expressing a preference for any country in future bids– this statistic is helpful for deciding what country is selected to organise SIGIR?

4 ...and together *with whom* did they write?

At the SIGIR 2006 conference in Seattle, Jon Kleinberg gave a keynote speech about *social networks, incentives and search* [3]. Kleinberg discussed one of the best known studies of social networks by Jeffery Travers and Stanley Milgram in the late 1960s [6]. Their findings are known as *six degrees of separation*. Travers and Milgram asked randomly chosen starters to forward a letter to a designated target individual. However, they could only send a letter to someone they knew on a "first-name basis". The study showed that many pairs of people were connected via short paths. The average path length was only six, hence "six degrees of separation".

In our analysis of the SIGIR proceedings we built a social network by using co-authorship as an equivalent of "knowing someone on a first-name basis". Suppose we take the SIGIR authors as nodes of a graph: Each node in the graph corresponds to a unique author and two nodes are connected if their corresponding authors have written a SIGIR paper together. By 2007, 1622 different authors have written a total of 1150 full papers. Poster papers were not included in the study. Figure 4 shows that, while in the early years the number of papers and authors were very much keeping up, in recent years a gap opened due to a significant increase in the average number of authors per paper. Apparently the number of routiniers beats the number of newcomers.

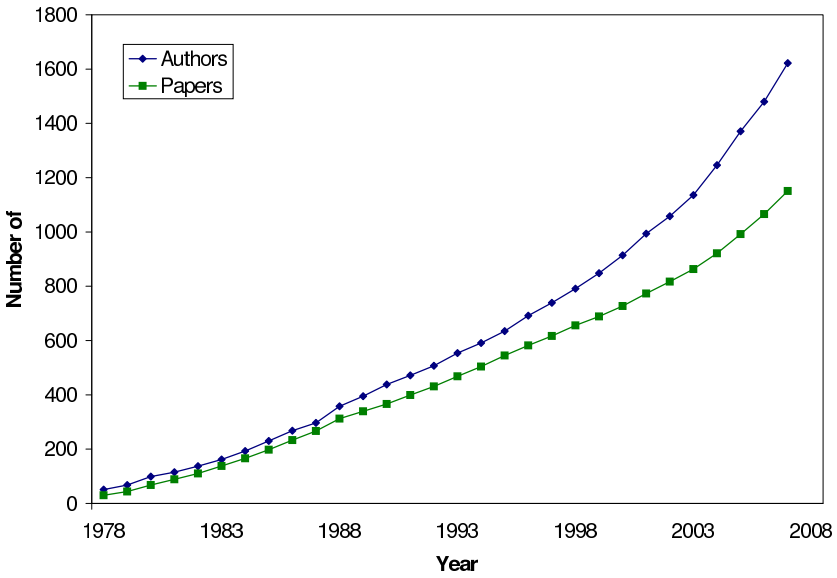

Figure 4: Development of the total number of authors and papers

The co-author graph consists of 361 different components, with one main component containing 635 SIGIR authors. The average path length in the largest component of 635 connected SIGIR authors is slightly higher than the six reported in the experiments of

Travers and Milgram. The complete graph is at: `http://pathfinder.cs.utwente.nl/sigir/images/big-graph.png`.

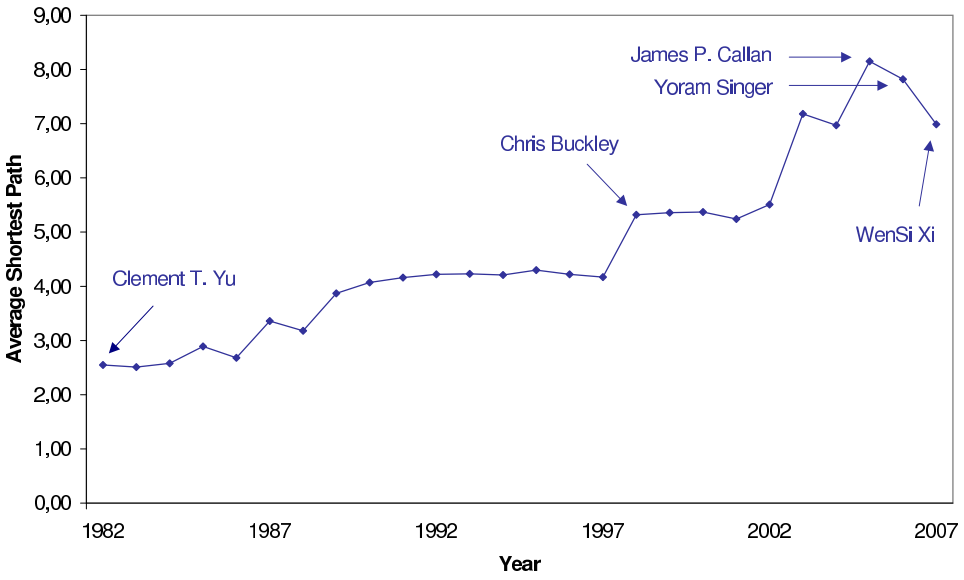

Figure 5: Development of the shortest average path length.

In order to determine the center of the graph [4, 5], we calculated for each of the authors in the largest component their distances to all other authors. The average distance of one author to all other authors is a measure of the centrality of the author in the graphs. The person with the shortest average path length is the center of SIGIR. To become the center of the graph it is not important how much you publish but with whom. In 2007 the center of SIGIR is Wensi Xi, closely followed by Bruce Croft and Edward Fox. The centers of the last 30 years are shown in Figure 5.

author	#co-authors
Wei-Ying Ma	54
W. Bruce Croft	41
Zheng Chen	36
James P. Callan	28
Clement T. Yu	26

author	#papers
W. Bruce Croft	44
James P. Callan	21
Wei-Ying Ma	18
James Allan	16
ChengXiang Zhai	16

Table 2: a) Top collaborating authors b) Top writing authors

Two further measures that are of interest are the number of papers written and the number of collaborations (Table 2). By far the most active author is Bruce Croft with more than double the number of papers in comparison to James Callan, the second in row. The author with the most co-authors is Wei-Ying Ma, closely followed by Bruce Croft.

5 Conclusion

From the patterns described above, some stereotypical paper profiles emerge. If you would like to write a paper for SIGIR that would count as nostalgic, you could write a paper about *"Computers for bibliographic data base records"*. In case you prefer to write a trendy paper for SIGIR, it is advisable to write about *"Our cross-language latent semantic web text question answering classification"*, find some co-authors in Singapore, preferably some that are central authors in the graph – which puts you close to the centre – and some that wrote SIGIR papers before but are currently not connected to the main graph – making you the one that connects them to all other authors, and possibly the next centre of the SIGIR universe.

Acknowledgements

This work was funded in part by the Dutch government research programme Multimedian.

References

[1] Franciska de Jong, Henning Rode, and Djoerd Hiemstra. Temporal language models for the disclosure of historical text. In *Proceedings of the 16th International Conference of the Association for History and Computing (AHC'05)*, pages 161–168, 2005.

[2] Djoerd Hiemstra, Henning Rode, Roel van Os, and Jan Flokstra. PF/Tijah: text search in an XML database system. In *Proceedings of the 2nd International Workshop on Open Source Information Retrieval (OSIR)*, 2006. http://dbappl.cs.utwente.nl/pftijah.

[3] Jon Kleinberg. Social networks, incentives, and search. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 210–211, 2006.

[4] Mario A. Nascimento, Jörg Sander, and Jeffrey Pound. Analysis of SIGMOD's coauthorship graph. *SIGMOD Record*, 32(3), 2003.

[5] Alan F. Smeaton, Gary Keogh, Cathal Gurrin, Kieran McDonald, and Tom Sodring. Analysis of papers from twenty-five years of SIGIR conferences: What have we been doing for the last quarter of a century? *SIGIR Forum*, 36(2), 2002.

[6] Jeffery Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969.