Analysis of Papers from Twenty-Five Years of SIGIR Conferences: What Have We Been Doing for the Last Quarter of a Century ?

Alan F. Smeaton, Gary Keogh, Cathal Gurrin, Kieran McDonald and Tom Sødring
Dublin City University, Glasnevin, Dublin 9, IRELAND
Alan.Smeaton@dcu.ie

As part of the celebration of twenty-five years of ACM SIGIR conferences we performed a content analysis of all papers published in the proceedings of SIGIR conferences, including those from 2002. From this we determined, using information retrieval approaches of course, which topics had come and gone over the last two and a half decades, and which topics are currently "hot". We also performed a co-authorship analysis among authors of the 853 SIGIR conference papers to determine which author is the most "central" in terms of a co-authorship graph and is our equivalent of Paul Erdös in Mathematics. In the first section we report on the content analysis, leading to our prediction as to the most topical paper likely to appear at SIGIR2003. In the second section we present details of our co-authorship analysis, revealing who is the "Christopher Lee" of SIGIR, and in the final section we give pointers to where readers who are SIGIR conference paper authors may find details of where they fit into the co-authorship graph.

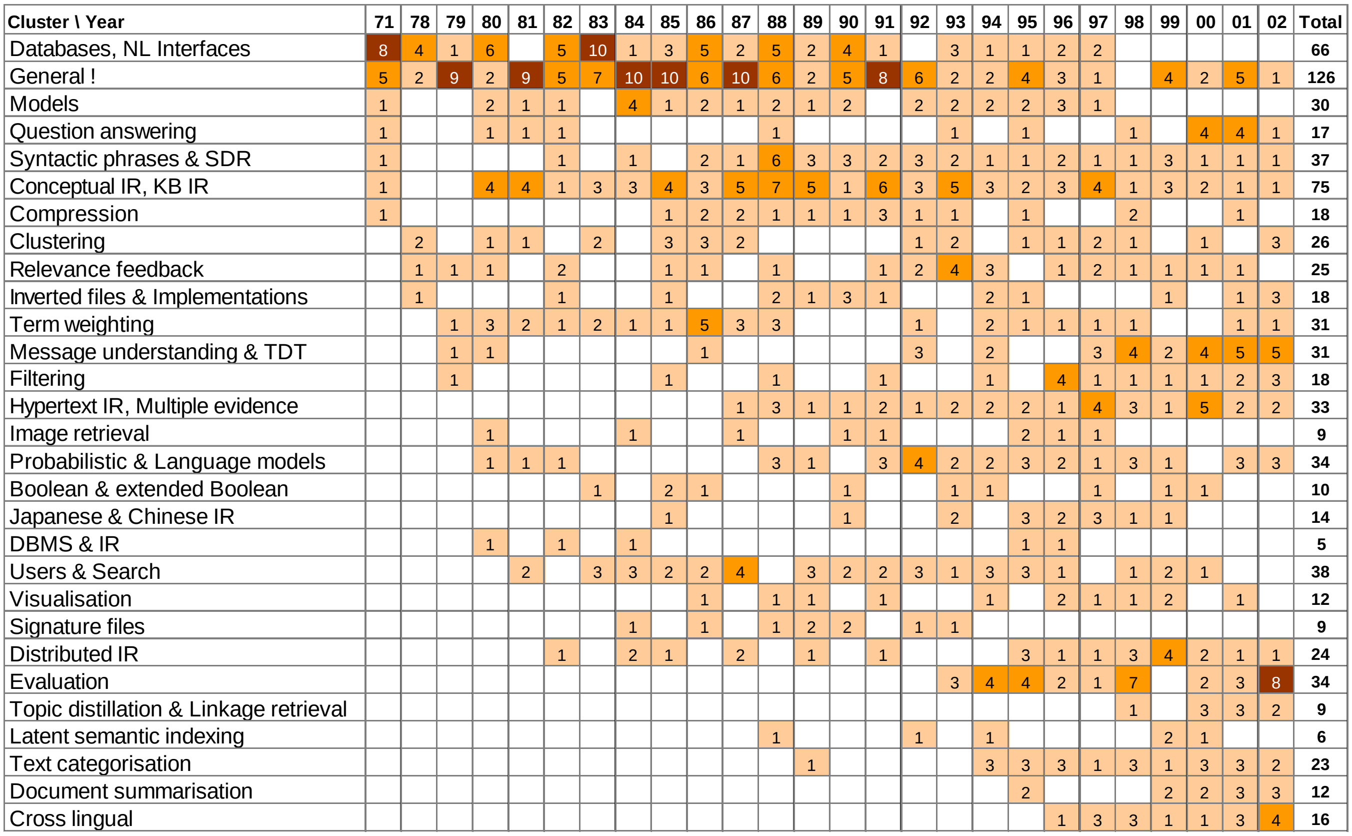
Content Analysis of SIGIR Conference Papers

In order to determine what topic areas are appearing in the papers at the SIGIR conferences from the last 25 years and to map those trends, we obtained the title, authornames, abstracts and year of publication of all 853 papers published.[1] We then applied Porter stemming and stopword removal to this text, represented terms from the title fields with twice the weights of author or abstract fields, and weighted each term using BM25 term weighting. Finally, we calculated an 853x853 similarity matrix for this set of documents and used Clustan Graphics version 5.25 [1] to generate an hierarchical, non-overlapping clustering of the document set.

We chose to use Clustan Graphics because it has a very user-friendly interface which allows a full-screen visualisation of the hierarchical clustering and allows the user to run a slider across the screen, effectively varying the similarity threshold above which clusters are created. This means that by using this slider, the user can not only see how many clusters are generated, but also how large these clusters are relative to each other. In our case we wanted to generate a number of clusters where the variability in size was small. We eventually settled on a threshold value which yielded 29 clusters, the smallest of which had only 5 documents, and the largest of which had 126. We then inspected each cluster manually, and assigned a topic description to reflect the theme of the majority of the papers in each cluster, which is a bit subjective but is as best we could.

[1] Some of these materials came from the ACM Digital Library and others were provided by OCR of PDF files. Many thanks to Jamie Callan for making this happen. The final, missing proceedings from the early 1980s came from Keith van Rijsbergen.

This largest cluster of 126 papers was larger than what we wanted but the topic variability in this cluster was very large, and effectively it was a "dumping ground" for papers whose similarities to others in the collection was small, and we regarded these as outliers. To add some structure to this clustering, and see how topics were spread over the last 25 SIGIR conferences, we mapped the documents in each cluster to the year of the SIGIR in which they appeared, and this is summarised in Figure 1.



Cluster \ Year	71	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	Total
Databases, NL Interfaces	8	4	1	6		5	10	1	3	5	2	5	2	4	1		3	1	1	2	2						66
General !	5	2	9	2	9	5	7	10	10	6	10	6	2	5	8	6	2	2	4	3	1		4	2	5	1	126
Models	1			2	1	1		4	1	2	1	2	1	2		2	2	2	2	3	1						30
Question answering	1			1	1	1						1					1		1			1		4	4	1	17
Syntactic phrases & SDR	1					1		1		2	1	6	3	3	2	3	2	1	1	2	1	1	3	1	1	1	37
Conceptual IR, KB IR	1			4	4	1	3	3	4	3	5	7	5	1	6	3	5	3	2	3	4	1	3	2	1	1	75
Compression	1								1	2	2	1	1	1	3	1	1		1			2			1		18
Clustering		2		1	1		2		3	3	2					1	2		1	1	2	1		1		3	26
Relevance feedback		1	1	1		2			1	1		1			1	2	4	3		1	2	1	1	1	1		25
Inverted files & Implementations		1				1			1			2	1	3	1			2	1				1		1	3	18
Term weighting			1	3	2	1	2	1	1	5	3	3				1		2	1	1	1	1			1	1	31
Message understanding & TDT			1	1						1						3		2			3	4	2	4	5	5	31
Filtering			1						1			1			1			1		4	1	1	1	1	2	3	18
Hypertext IR, Multiple evidence											1	3	1	1	2	1	2	2	2	1	4	3	1	5	2	2	33
Image retrieval				1				1			1			1	1				2	1	1						9
Probabilistic & Language models				1	1	1						3	1		3	4	2	2	3	2	1	3	1		3	3	34
Boolean & extended Boolean							1		2	1				1			1	1			1		1	1			10
Japanese & Chinese IR									1					1			2		3	2	3	1	1				14
DBMS & IR				1		1		1											1	1							5
Users & Search					2		3	3	2	2	4		3	2	2	3	1	3	3	1		1	2	1			38
Visualisation										1		1	1		1			1		2	1	1	2		1		12
Signature files								1		1		1	2	2		1	1										9
Distributed IR						1		2	1		2		1		1				3	1	1	3	4	2	1	1	24
Evaluation																	3	4	4	2	1	7		2	3	8	34
Topic distillation & Linkage retrieval																						1		3	3	2	9
Latent semantic indexing												1				1		1					2	1			6
Text categorisation													1					3	3	3	1	3	1	3	3	2	23
Document summarisation																			2				2	2	3	3	12
Cross lingual																				1	3	3	1	1	3	4	16

Figure 1: Clustering of SIGIR papers by topic vs. year

In Figure 1 the rows in the table, representing clusters or topics, are sorted approximately in order of a combination of the year of their first appearance, and the number of papers published. Each cell in the table is coloured to help visualise the number of documents present. In the first row we can see that the topic " Databases, NL Interfaces" (which is actually two topics) appeared strongly in the 1980s and has diminished as we move into the 1990s and the present decade. We might have expected "question answering" to be a recent development but we can see from row 4 that QA has actually been ongoing since the first SIGIR in 1971. "Cross lingual IR" represented in the last row, first appeared in 1996 and has been growing in representation since while the growing presence of "Evaluation" from 1993 is nicely correlated with the growing impact of TREC, which started in 1991.

SIGIR Co-Authorship

With a machine-readable version of the SIGIR authors available we were able to manually clean up author names (e.g. Alan Smeaton = A.F. Smeaton) and create a co-author collaboration graph. This allowed us to explore SIGIR co-authorship and calculate some of the easy and obvious things. For example, the author with the greatest number of SIGIR papers is Bruce Croft, and the author with the greatest number of collaborators is also Bruce, with 31, followed by Jamie Callan with 22 and Clement Yu with 19. However, much more interesting to explore are the paths between authors on this graph, and so we decided to apply Erdös-type analysis to this.

Mathematical folklore has evolved around Paul Erdös and his co-authorships. Erdös was a Hungarian mathematician who had an enormous number of scientific papers (at least 1401) and a huge number of co-author collaborators (at least 502). Partly out of respect for this wonderful individual, and as a mathematical challenge, the mathematical community has created a collaboration graph for its community with approximately 337,000 authors of 1.6 million authored items in the Math Review database, and Paul Erdös is in the centre of that graph. An "Erdös number" is the smallest number of co-authorship links between an individual and Paul Erdös. One of the authors of this paper has an Erdös number of 4 (Gary Keogh), as has Susan Dumais (along with 62,134 other individuals), but Alistair Moffat has an Erdös number of only 2 ! An extensive website exists which allows people to look up their own Erdös number and has much more related information [2] and the whole concept of Erdös number distances is still a topic for research in the field of scientometrics [3].

The collaboration graph among SIGIR authors is very fragmented, with one large and several smaller components, as well as many disconnected authors or author pairs. The largest components are 211, 48, 26, 20, 14, 13, 13, 12, and so on, in size. There are some surprises in the collaboration graph, for example the fact that Sue Dumais (current SIGIR chair) and Bruce Croft (previous SIGIR chair) have a distance of 7, with Bruce Croft to Victor Lavrenko to James Allan to Chris Buckley to Clement Yu to Abraham Bookstein to Scott Deerwester to Sue Dumais. Given that Sue has an Erdös number of 4, that means Bruce has an Erdös number of at most 11 !

Pairwise collaboration graphs like this have been constructed for many things including chess matches and musicians in rock bands but possibly the most famous of these is the Oracle of Kevin Bacon. This covers 512,126 actors who have acted together in more than 275,000 movies, and is updated regularly from the Internet Movie database. This collaboration graph is much more tightly connected than the graph of SIGIR paper authors and the "oracle" or centre is the actor or actress who has the shortest average path length to all other actors or actresses in the graph. When this was being created it was hypothesised that it would be Kevin Bacon, and actor who has appeared in many movies, far too many of which are duds, but as the collaboration graph is updated regularly, this centre of the Hollywood Universe will change. At the time of doing our SIGIR analysis the centre was Christopher Lee (avg. path length 2.622940) followed by Rod Steiger and Donald Pleasence.

So, if the centre of mathematics is Paul Erdös, and the centre of Hollywood is Christopher Lee, who is the centre of SIGIR ? Clearly, our centre should come from the cluster of 211 authors, and should be the ACM SIGIR Christopher Lee Award for the SIGIR Conference author most closely connected to the greatest number of other authors in terms of co-authorship links. The announcement of this was made at the SIGIR2002 conference dinner and is Chris Buckley (path length 3.65) followed by Gerry Salton (3.76), James Allan (3.791), Clement Yu (3.862) and Amit Singhal (3.895). To mark the occasion, Chris was presented with a certificate, and a Christopher Lee DVD (Lord of the Rings).

Inevitably, on carrying out the co-authorship exercise, we discovered many things we would like to do to extend this work. We would like to make the co-authorship graph weighted to take account of more than paper co-authored by a pair of authors. We would also like to normalise links to factor in the number of authors of a paper, allowing for a stronger link between authors where there are only two authors of a paper, and a weaker link where there are more than two authors. We would also like to broaden the source of co-author links and include other relevant conferences that SIGIR co-sponsors (CIKM, ACM DL, JCDL, etc.) and include relevant journals (J.ASIS, IP&M, IRJ, ACM TOIS, etc.). All of these ideas, and others, represent a wish list which we may or may not get to do.

What is the Ideal SIGIR2003 Paper and where are you on the Graph ?

So what has this analysis taught us ? Apart from attempting to track the evolution of topic areas in our field, which is interesting, we can also extrapolate and predict the "hottest" topics for next year. From the co-authorship graph, we can also prescribe the co-authorship combination which goes furthest to "unite" our graph. Thus the ideal paper to appear in SIGIR2003 should be titled "Evaluation of a Language Model Implementation of a Topic-Based, Cross-Lingual Question-Answering and Summarisation System", and it should be by Chris Buckley, Keith van Rijsbergen, and Jian-Yun Nie.[2]

Readers of this paper who are also authors of past SIGIR conference papers might like to see where they fit into the co-authorship graph and calculate their own "Buckley number". To do so we have created a WWW interface to our database allowing a user to see how "far" they are from Chris Buckley, or to calculate the distance between any two SIGIR conference paper authors. The URL for this is http://www.cdvp.dcu.ie/SIGIR/ where there are links to slides of the presentation at SIGIR2002, and other resources.

Acknowledgement

Funding for a summer student to carry out much of this work was provided by ACM SIGIR, for which we are grateful.

[2] The fact that one of the authors of this paper is also one of the program co-chairs of SIGIR2003 is purely co-incidental and no bias in favour of such a paper submitted to SIGIR2003 will be given ;-)

References

[1] Clustan Graphics available at http://www.clustan.com/

[2] The Erdös Number Page available at
http://www.oakland.edu/~grossman/erdoshp.html

[3] Co-authorship, Rational Erdös Numbers and Resistance Distances in Graphs. A.T Balaban and D.J. Klein, *Scientometrics*, 55(1), (2002), 59-70.