Institution: City University London

Unit of Assessment: 11 Computer Science and Informatics

Title of case study: Making the results returned by search engines more relevant

1. Summary of the impact

The user experience of searching the web is usually a very positive one, in part due to the work carried out at City University London on obtaining more relevant documents on the first page of search results. The model produced in our work outperforms other methods in benchmark tests and helps users to access better quality information billions of times every day. Evidence from a variety of sources shows that the work has had a significant economic impact nationally and internationally. Many software companies have benefited from the work, including multinationals (Microsoft) and UK small and medium-sized enterprises (SMEs) (Grapeshot) and those who use the services of such software, including Reed Recruitment, MyDeco and UNESCO. Getting the right information to people efficiently and reducing the number of searches performed saves time and money and has a wide range of benefits for individuals and society.

2. Underpinning research

A core requirement of information retrieval (IR) systems is to obtain as many relevant documents for the user as possible in any given search. A key part of meeting this requirement is ranking information according to the relevance of the retrieved items to satisfy the user's information need. This work focuses on ranking text documents given their relevance to the user, using document length information in conjunction with other statistics such as word frequency – this is known as a 'ranking function'. It is based on a strong theoretical model in probability theory, the Robertson/*Spärck* Jones probabilistic retrieval model. This ranking function is known as BM25 – BM for 'Best Match', 25 for the function number in software. The work was carried out by Stephen Robertson, Professor of Information Science (full time September 1978 to April 1998, part time April 1998 to December 2009, Emeritus since January 2010), who undertook the theoretical work, together with Stephen Walker, Research Fellow (1988–1998), who implemented the model in software.

The BM25 ranking function was tested at an annual conference run in the USA, the Text REtrieval Conference (TREC). This is a competition that allows participants to compare the results from a given IR task and has become central to research in the field of search. The breakthrough for the model was at the 1994 conference (the paper in which these results were published has been cited over 1,000 times[2]). The team was able to build on work in the previous two conferences to demonstrate for the first time that BM25 could significantly improve retrieval results on the same dataset compared with other participants.

The team tested the ideas in two main tasks: *ad hoc* and routing[10]. A total of 33 groups, 14 companies and 19 universities, participated in the competition. The *ad hoc* task is a normal Google search. The routing task is a filtering application, with documents selected to satisfy a user profile based on prior judgements of relevance. In both tracks, the methods outperformed other all other methods embodying a variety of different systems and models (e.g., all other groups participating in the tracks). The results was that the BM25 matching function became the baseline against which all other systems and models compare themselves. Few if any systems have been able to provide better retrieval results than the model developed in this seminal work[6], and the BM25 function remains one of the most widely used baseline functions in IR research.

Further development of the model to incorporate additional evidence was carried out in collaboration with Microsoft Research Cambridge, UK, with further development of the ranking function that focused on field weighting[4], which allows individual components of a document (title, abstract, etc.) to be weighted individually rather than weighting the whole document. Research using the model continues in the Department, for example Andrew MacFarlane and colleagues'

work on the optimisation of filtering/routing queries[1].

3. References to the research

The outputs listed below underwent rigorous peer review prior to acceptance for publication in a journal or are published in proceedings from conferences that are very highly regarded in the field.

1. MacFarlane A., Secker A., May P. & Timmis, J. (2010). An experimental comparison of a genetic algorithm and a hill-climber for term selection. *Journal of Documentation*, 66(4), 513–531 10.1108/00220411011052939

2. Robertson S. E., Walker S., Jones S., Hancock-Beaulieu M.M. & Gatford M. (1994). OKAPI at TREC-3. In D. Harman (Ed.), *NIST Special Publication 500-226: Overview of the Third Text REtrieval Conference (TREC-3)* 109–126 http://trec.nist.gov/pubs/trec3/t3_proceedings.html [Citations: 1066]

3. Robertson S.E. (1997). Overview of the Okapi projects. Special issue of *Journal of Documentation*, 53(1), 3–7 10.1108/EUM0000000007186 [Citations: 164].

4. Taylor M., Zaragoza H., Craswell N., Robertson S.E. & Burges, C. (2007). Optimisation methods for ranking functions with multiple parameters. In *CIKM '06 Proceedings of the 15th ACM International Conference on Information and Knowledge Management* 585–593 10.1145/1183614.1183698

4. Details of the impact

The TREC collections and methodologies have been established in the past 20 years as the *de facto* standard with which IR researchers publish results that are defensible, comparable and reproducible[12]. The 2010 Rowe survey of the economic impact of TREC on the field, particularly in the context of the USA, found that the impact was significant and that TREC has been successful for companies producing IR software and search services. Rowe *et al*. also state:

> *"TREC has made significant contributions to the technology infrastructure supporting IR system development, the benefits of which flow directly or indirectly to a variety of stakeholder groups…. The direct beneficiaries are IR researchers in academic research groups and commercial firms; TREC's accomplishments improved both the efficiency and the effectiveness of their research and development (R&D) activities. R&D benefits that accrued to academic labs have also flowed indirectly to commercial firms through technology transfer and knowledge sharing. Improvement in the R&D of commercial IR firms led to improvements in the performance of IR systems commercialized into products and services. End users of these IR systems have also indirectly benefited from TREC through higher quality IR products and Services."*

Armstrong *et al*.[6] reference BM25 as the best system from TREC 3, 'which remains one of the best systems in the entire 12 year dataset'. Further evidence of the impact of BM25 comes from implementation of the model in IR software and services. Although it is not always clear what algorithms are used by commercial enterprises, many of the participants of the TREC conference are commercial organisations and the BM25 model has been influential. For example, Microsoft now use a form of the ranking function in both their Bing search engine (first introduced in 2005), and their SharePoint enterprise search (first introduced in 2003). Bing is now the second largest web search engine after Google, having overtaken Yahoo in December 2011.

The model has been successfully implemented in a variety of ways by Dr Martin Porter (Technical Director) for Grapeshot Ltd to provide better search services for clients. Grapeshot is a UK SME that uses advanced IR techniques to assess the relevant significance of keywords in pages and what users read, to support online advertising placement. The following statement on the impact of BM25 has been provided by John Snyder, CEO of GrapeShot:

> *"I am indeed very proud to say as CEO of Grapeshot, where we employ over 20 highly technical and clever people, that we have productized the BM25 work into a suite of software web services that make search calculations over 12 billion times per month.*
>
> *Our Grapeshot software, with your probabilistic information retrieval work at its heart, is used by a majority of UK publishers such as Mail Online, Telegraph, Independent, Mirror, Johnston Press, Reuters, IPC Media, Future Publishing, Incisive Media to help them make more revenues from targeted online advertising. In essence the publisher page, in real-time, seeks the best advertising to be contextually placed on the page. So this is probabilistic information retrieval working in milliseconds, many thousand times a second.*
>
> *We do serve international customers such as Glam Media and Verisign in the USA, but the majority of our products are used by UK customers, and all our staff work in Cambridge or London at our development or sales offices."*

This passage refers to advertising search revenue, which is a global multi-billion dollar commercial activity. In the first quarter of 2012, the total global revenue for online advertising was $8.4B[11]. By being implemented on search engines such as Bing, the BM25 model has had a significant economic impact globally as better ranking leads to results sets with more relevant documents, a high user acceptance and therefore greater advertising revenues.

The model has also been implemented in widely used open source software packages including Apache Lucene (Solr, Cassandra), Xapian and Greenstone. Apache Lucene is a widely used IR library, which is an integral part of the highly regarded Apache Software Foundation open source software projects. Software from this project was used on 65.4% of websites in September 2013. Flax, an enterprise search engine built on Xapian, has been used to provide search infrastructure for companies such as the Government Digital Service, Newspaper Licensing Agency, TMC Marine, C Spence Ltd, Australian Associated Press, Reed Specialist Recruitment, *Financial Times*, Durrants and MyDeco.

The Government Digital Service is a unit within the UK Cabinet Office tasked with transforming Government digital services to ensure that Government offers world-class digital products which meet people's needs. This includes ensuring appropriate forms of support for people who are unable to access or use digital services, and developing Gov.uk, the single domain for Government, making it simpler, clearer and faster to access government information and services. Reed Specialist Recruitment (part of Reed Global) significantly improved the search for various stakeholders such as jobseekers and recruiters to more efficiently bring these two together, identifying the right person for the right job. MyDeco provides searches on various types of product to enable more efficient e-commerce for the purchase of items such as furniture and fittings. The Xapian library is used in Debian distributions, 139,988 having installed the package by September 2013. Greenstone (partly developed at City University London) is a widely used open source digital library system, which is supported by various United Nations organisations such as the United Nations Educational Scientific and Cultural Organization (UNESCO). UNESCO actively encourages the use of Greenstone for all of its activities, including education, natural sciences and social sciences, and provides support for client organisations. This impact is significant in terms of economics and to support cultural heritage preservation programmes.

There is plenty of evidence from a variety of sources that the BM25 matching function has had a significant economic impact nationally and internationally. The reach of the work is wide-ranging, benefiting many software companies, including multinationals (Microsoft) and UK SMEs (Grapeshot), and users of such software, for example the *Financial Times* and UNESCO.

5. Sources to corroborate the impact

5. http://search-lucene.com/jd/lucene/core/org/apache/lucene/search/similarities/BM25Similarity.html#BM25Similarity%28%29.

6. Armstrong, T., Moffat, A., Webber, W., & Zobel, J. (2009). Has adhoc retrieval improved since 1994? In J. Allan, J. Aslam, M. Sanderson, C. X. Zhai, & Zobel, J. (Eds.), *Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval*, Boston, Massachussetts, July 2009 (pp. 692–693). http://www.csse.unimelb.edu.au/~jz/fulltext/sigir09.pdf.

7. Craswell, N. (2012). Confidential email correspondence with Professor S. Robertson in connection with use of BM25 by Microsoft in their products; available on request from City.

8. www.flax.co.uk/blog/2009/04/02/xapian-search-architecture/ ; www.flax.co.uk/downloads/ (case studies).

9. www.greenstone.org/ http://tinyurl.com/k53y9gz (UNESCO Greenstone page).

10. Harman, D. (1994). Overview of the third Text REtrieval Conference. In D. Harman (Ed.), NIST Special Publication 500-226: *Overview of the Third Text REtrieval Conference (TREC-3)* (pp. 1–20). http://trec.nist.gov/pubs/trec3/t3_proceedings.html.

11. BusinessWire (2012). Internet advertising revenues set first quarter record at $8.4 billion. www.businesswire.com/news/home/20120611005230/en/Internet-Advertising-Revenues-Set-Quarter-Record-8.4.

12. Rowe, B. R., Wood, D. W., Link, A. N., & Simoni, D. A. (2010). Economic impact assessment of NIST's Text REtrieval Conference (TREC) Program. RTI Project Number 0211875. http://trec.nist.gov/pubs/2010.economic.impact.pdf.

13. Snyder, J. (2012). Confidential statement on impact of BM25 on Grapeshot; available on request from City.

14. http://xapian.org/docs/bm25.html.