ECIR 2008 Workshop on Efficiency Issues on Information Retrieval

Roi Blanco
IRLab, University of A Coruña
Spain
rblanco@udc.es

Fabrizio Silvestri
ISTI-CNR, Pisa
Italy
fabrizio.silvestri@isti.cnr.it

1 Introduction

The goal of EIIR 2008, the first Workshop on Efficiency Issues in Information Retrieval, was to shed light on efficiency-related issues of modern high-scale information retrieval (IR), e.g., Web, distributed technologies, peer to peer architectures and also new IR environments such as desktop search, enterprise/expert search, mobile devices, etc. In addition, the workshop aimed at fostering collaboration between different research groups in this area.

The vast amounts of digitally available information create the need for systems that retrieve information in an effective and efficient way. IR effectiveness deals with retrieving the most relevant information to a user need, while IR efficiency deals with providing fast and ordered access to large amounts of information. IR efficiency ensures that systems scale up to the vast amounts of information available for retrieval, and is of utmost importance to both academic and corporate environments. In academia, it is imperative for new ideas and techniques to be evaluated on as near-realistic environments as possible; this is reflected in the past Terabyte track and recent Million Query track organised by the Text REtrieval Evaluation Conferences (TREC). In corporate environments, it is important that systems response time is kept low, and the amount of data processed high. These efficiency concerns need to be addressed in a principled way, so that they can be adapted to new platforms and environments, such as IR from mobile devices, desktop search, distributed peer to peer, expert search, multimedia retrieval, and so on. Efficiency research over the past years has focused on efficient indexing, storage (compression) and retrieval of data (query processing strategies).

The EIIR 2008 workshop was held on March 30, 2008 in conjunction with ECIR 2008. The workshop focused on efficiency-related problems of high-scale IR and new IR environments. Past workshops on this area include: the Workshop on Large-Scale Distributed Systems, collocated with SIGIR 2007; the Workshop on Information Retrieval in Peer-to-Peer Networks collocated with SIGIR 2004 and with CIKM 2005; and the Workshop on Heterogeneous Distributed Information Retrieval, collocated with SIGIR 2005. This year's workshop attracted around 12 researchers from both academia and companies (Google, Yahoo! and IBM), working on the efficiency of distributed, XML, multimedia, and general IR systems. The workshop format included an one-hour keynote speech, 30 minutes presentations for accepted papers, and one hour for a guided discussion. The workshop's friendly atmosphere allowed for a thorough exchange of ideas and useful discussions on every paper, hence helping the corresponding authors to sketch future lines of research.

The call for papers attracted submissions from within and outside Europe. The program committee accepted 6 scientific papers presenting original solutions for efficiency issues in different environments and tasks.

The workshop was opened by Diego Puppin, from Google, Boston, USA, who gave an invited talk on efficiency trade-offs of modern IR systems. The keynote talk provoked a lively discussion that continued throughout the paper presentations and the closing discussion. Paper presentations focused on efficient indexing, compression and matching, as well as efficient distributed XML, collaborative filtering and image IR. The discussion was organised around the usage scenarios and economic factors steering the adoption of the efficient technology in the field of IR.

The workshop program, talks and slides are available on-line from http://irlab.dc.fi.udc.es/ecir/.

2 Presentation summaries

The invited talk given by *Diego Puppin* on **The Index, the Cache and the Company's Cash: Daily Trade-offs for the Modern Information Retriever** motivated the need for controlling the resources and cash put into crawling, indexing, querying and serving IR results. The talk had two parts. First, Diego commented on some of the efficiency concerns that a modern scalable Web search engine must cope with, in order to deal with billions of daily queries, focusing on the particular case of Google, and gave many insightful real-life examples from a leading company in the field. In the second part of the talk, Diego gave an overview of existing distributed search infrastructures, such as the document-partitioned and the pipelined term-partitioned architectures, and commented on their limitations. A new approach was proposed for reducing the number of servers utilized by each query, by learning from users behaviour, and representing documents as 'bags of queries'. The proposed technique was able to guarantee, for repeated queries, results comparable to those obtained with a centralized index, at a fraction of the cost.

The first presentation was given by *Judith Winter* on **A Distributed Indexing Strategy for Efficient XML Retrieval**, which was collaborative work with Oswald Drobnik, from J.W.Goethe-University, Frankfurt. Judith presented a novel indexing strategy for XML IR in P2P systems, which indexes documents either globally into distributed indexes or locally in connection with distributing peer summaries, depending on a peer's status. Then, structural information from XML documents is used to select entries for pruned posting and peer lists.

Fidel Cacheda presented **Algorithms for Efficient Collaborative Filtering**, which was work realised with Vreixo Formoso and Victor Carneiro, from University of A Coruña. A series of collaborative filtering algorithms known for their simplicity and efficiency was presented (namely item mean, simple mean-based, and tendencies-based). The efficiency of these algorithms was compared with that of other well-known high-performing collaborative filtering algorithms. The proposed algorithms had better response times (at least by two orders of magnitude), in the training as well as when making predictions. Experimental results allowed to conclude that, despite of the fact that the algorithms are significantly more efficient than those proposed in previous work, they perform comparably to state-of-the-art techniques for collaborative filtering.

Srikanta Bedathur presented **Tunable Word-Level Index Compression for Versioned Corpora**, which was joint work with Klaus Berberich and Gerhard Weikum. Srikanta presented a tunable method for compressing indices of versioned corpora that supports time-travel phrase queries. This method provides phrase querying over evolving collections, as of a given time. Word positions, necessary in phrase querying, were fused in many neighbouring versions of a document, and the index was compressed accordingly, using the proposed FUSION compression scheme, which was also tunable w.r.t. query-processing overheads. In brief, the high level of redundancy between versions (i.e. positional proximity of terms in consecutive versions) is fully exploited by the compression mechanism. The compression technique reduces the final index size in more than 50% when compared to the baseline, with a query-processing overhead of less than the 50% in the worst case.

Patrice Lacour presented **Efficiency Comparison of Document Matching Techniques**, which was joint work with Craig Macdonald and Iadh Ounis from the University of Glasgow. Patrice reviewed several state-of-the-art approaches for matching documents to query terms, based on term-centric and document-centric scoring, using three modern Web IR test collections, and concluded in terms of the trade-off between retrieval effectiveness and efficiency, using uniform settings, the same retrieval platform, and on several collections.

The last two papers focused on efficient image retrieval. *David García-Pérez* presented **Evaluation of a M-Tree in a Content-Based Image Retrieval System**, which was collaborative work with Antonio Mosquera (University of Santiago de Compostela), Stefano Berretti and Alberto del Bimbo (University of Firenze). David proposed an efficient M-Tree index structure for an image feature extraction system which works with active nets. This structure speeds up the search process and reduces the necessary I/O to disk, when extracting high-dimensionality features from images. The M-Tree supports different graph-insertion schemes during its construction, which are needed for indexing those multi-dimensional features. Experimental results on a standard object database proved that the M-Tree speeds up the retrieval process significantly.

The last presentation was given by *Shai Erera* (IBM Haifa) on **Metric Inverted - an Efficient Inverted Indexing Method for Metric Spaces**, which was a work by Benjamin Sznajder, Jonathan Mamou, Yosi Mass and Michal Shmueli-Scheuer, from IBM Lab, Haifa. Authors proposed a framework for efficient indexing and retrieval of audio-visual content by extending classical techniques taken from the textual IR methods such as lexicon, posting lists and boolean constraints. The motivation for this work is that multimedia retrieval is still limited to manually added metadata, and that state of the art solutions for content based image retrieval do not scale. The main claim is that by reducing the gap between textual and multimedia retrieval (using lexicons of features, and indexing objects as canonical forms of the features found in the lexicon) it is possible to develop methods usable in real Web environments. The scalability of the proposed method was experimentally evaluated on a large image collection, and the results are very promising, as the method proves to be very effective and improves drastically the efficiency (about 90%) over state-of-the-art methods.

3 Discussion

The discussion began with the question *Aren't commercial search engines enough?* [1] Major large-scale search engines possess user data and computer hardware that researches cannot begin to reproduce. These companies have changed the way people find information and also shifted the balance of knowledge between industry and academia. This is especially important if we focus on efficiency of systems and algorithms, where the problem of scale is the core of research.

There was common agreement on two major points. The first point was that major commercial search engines are not going to address and solve every problem in the field, and also they can be reluctant to disclose figures or techniques dealing with efficiency issues. The second point was that the recent availability of open source systems has helped enormously to provide technological background and baselines. Someone suggested that academic research results from an efficiency point of view in Web IR may be unproductive as they may not be really scalable, or entirely representative of real systems. Re-creating or simulating those conditions at a smaller scale was suggested as an alternative, but received mixed feedback. Another controversial suggestion about Web IR efficiency was that it is becoming increasingly impossible to provide efficiency figures, with the exception of Google or Yahoo!, and that perhaps it would be better to move on to other environments, like enterprise search for example. Even though this remark was acknowledged, it was nevertheless pointed out that most work on IR efficiency has been done within academic groups. The discussion was closed with a proposal to make available some more data from commercial companies to practitioners.

4 Conclusions

EIIR was a successful event which brought together researchers and practitioners from the academia and industry, with high quality presentations and lively discussions, that have stirred ideas for on-going work.

We thank the authors and presenters, the invited speaker and the members of the program committee, and all workshop participants for their contributions to the presented material and productive discussions. This workshop has been partially funded by: “Rede Galega de Procesamento da Linguaxe e Recuperacion de Informacion” (Galician Network for Language Processing and Information Retrieval), funded by ‘Xunta de Galicia.

References

[1] Jamie Callan, James Allan, Charles L. A. Clarke, Susan Dumais, David A. Evans, Mark Sanderson and ChengXiang Zhai. Meeting of the MINDS: an information retrieval research agenda. In *ACM SIGIR Forum, Vol. 4, No. 2*, 2007