Workshop on Large-Scale Distributed Systems for Information Retrieval

Sebastian Michel, Gleb Skobeltsyn
Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{sebastian.michel, gleb.skobeltsyn}@epfl.ch

Wai Gen Yee
Illinois Institute of Technology, USA
waigen@ir.iit.edu

Abstract

Due to the dramatically increasing amount of available data, effective and scalable solutions for data organization and search are essential. Distributed solutions naturally provide promising alternatives to standard centralized approaches. With the computational power of thousands or millions of computers in clusters or peer-to-peer systems, the challenges that arise are manifold, ranging from efficient resource discovery to issues in load balancing and distributed query processing.

The 2008 edition of the Workshop on Large-Scale Distributed Systems for Information Retrieval (LSDS-IR'08) provided a forum for researchers to discuss these problems and to define new directions for the work on Distributed Information Retrieval. The Workshop program featured research contributions in the areas of similarity search, resource selection, network organization schemes, issues of data quality, result ranking techniques and query routing algorithms.

1 Introduction

We have witnessed an exponential growth of the amount of Web content in the past twenty years since the beginning of the World Wide Web in 1989. It is no surprise that searching this ocean of data poses serious challenges in terms of quality and speed. Web search engines such as *Google* and *Yahoo!* rely upon large complex systems to be able to handle thousands of queries per second making Web search a basic service in many aspects of our everyday life.

However, despite the general success of Web search, the problem of efficient and effective searching in large-scale data repositories is far from being solved. For instance, improving accuracy is the current focus of research in Web search engines. The Web comprises various types of content in the form of textual documents, structured metadata, databases, maps, images, video, etc. Using specific properties of each type could potentially increase the quality of search, but requires non-trivial solutions to handle large volumes of data. Search quality

related research directions include processing of structured queries, relevance computation, data freshness, spam detection, duplicate page removal, language detection, *etc.*

The main challenge, however, may be *scalability* – the ability to cope with the growing amount of information and the increasing demand for the search service. A recent survey [2] envisions that the number of servers required by a search engine to keep up with the load in 2010 may be in the order of millions. This could be infeasible for the cluster-based architecture currently employed by commercial search engines. Thus, it is important to design a truly distributed large-scale system that enables fast and accurate search over very large amounts of content.

The 6th edition of the Workshop on Large-Scale Distributed Systems for Information Retrieval featured nine papers on emerging work in the areas of similarity search, resource selection, network organization schemes, issues of data quality, result ranking techniques and query routing algorithms. The Workshop was held on October 30, 2008 in conjunction with the ACM CIKM conference in Napa, California, USA. It continued the series of workshops on distributed and peer-to-peer IR systems collocated with either CIKM or SIGIR conferences in the past: LSDS-IR in 2007; the Workshop on Information Retrieval in Peer-to-Peer Networks (P2PIR) in 2004, 2005 and 2006; and the Workshop on Heterogeneous and Distributed Information Retrieval (HDIR) in 2005.

In the remainder of this report, we present the Workshop program committee and the program. We then make a few comments on the keynote and the presented research papers. We finish with some comments on lessons learned.

2 Program Committee

The program committee for the Workshop included:

- Karl Aberer, Ecole Polytechnique Fédérale de Lausanne, Switzerland;
- Ricardo Baeza-Yates, Yahoo! Research Barcelona, Spain;
- Roi Blanco, University of A Coruña, Spain;
- Abdur Chowdhury, Illinois Institute of Technology, Chicago, USA;
- Ophir Frieder, Illinois Institute of Technology, Chicago, USA;
- Norbert Fuhr, University of Duisburg-Essen, Germany;
- Flavio Junqueira, Yahoo! Research Barcelona, Spain;
- Wolfgang Nejdl, University of Hannover, Germany;
- Salvatore Orlando, University of Venice, Italy;
- Josiane Xavier Parreira, Max-Planck-Institut Informatik, Germany;
- Raffaele Perego, ISTI-CNR, Italy;
- Vassilis Plachouras, Yahoo! Research Barcelona, Spain;
- Diego Puppin, Google, USA;
- Martin Rajman, Ecole Polytechnique Fédérale de Lausanne, Switzerland;
- Fabrizio Silvestri, ISTI-CNR, Italy;
- Torsten Suel, Polytechnic University, USA;

- Peter Triantafillou, University of Patras, Greece;
- Christos Tryfonopoulos, Max-Planck-Institut Informatik, Germany;
- Michalis Vazirgiannis, Athens Univ. of Economics & Business, Greece;
- Ivana Podnar Žarko, University of Zagreb, Croatia;
- Justin Zobel, RMIT University, Australia.

3 Workshop Program

The Workshop started with a keynote talk given by Torsten Suel entitled: "Search Engine Architectures from Conventional to P2P" [11]. The keynote gave an excellent overview of the state of the art in large-scale distributed systems for Information Retrieval. In the continuation, there were nine paper presentations in two technical sessions including one industry talk. The presentations were:

1. Fabrizio Falchi, Claudio Lucchese, Salvatore Orlando, Raffaele Perego and Fausto Rabitti: "A Metric Cache for Similarity Search" [7];
2. Stanislav Barton, Vlastislav Dohnal, Jan Sedmidubsky and Pavel Zezula: "Building Self-Organized Image Retrieval Network" [3];
3. Hans Witschel: "Ranking Information Resources in Peer-to-Peer Text Retrieval: an Experimental Study" [13];
4. Christos Doulkeridis, Kjetil Nørvåg and Michalis Vazirgiannis: "Peer-to-Peer Similarity Search over Widely Distributed Document Collections" [6];
5. Ira Woodhead: "Robot Army: A Distributed System for the Casual Manipulation of Massive Data Sets" (Industry talk) [1];
6. Pascal Felber, Toan Luu, Martin Rajman and Étienne Rivière: "Managing Collaborative Feedback Information for Distributed Retrieval" [8];
7. Venkata S. Cherukuri, K. Selcuk Candan: "Propagation-Vectors for Trees (PVT): Concise yet Effective Summaries for Hierarchical Data and Trees" [4];
8. Judith Winter: "Routing of Structured Queries in Large-Scale Distributed Systems" [12];
9. Dongmei Jia: "Cost-Effective Spam Detection in P2P File-Sharing Systems" [9].

The first two papers deal with two different issues in image retrieval. Falchi *et al.* are concerned with the ability to find quickly results within a single node via cache management, avoiding secondary storage access. Barton *et al.* aim at creating a workload-based topology to improve recall and decrease cost when searching for images in a peer-to-peer network. Performance, in this case, improves with usage.

The third and fourth papers deal with text search in peer-to-peer networks. Witschel focuses on how peer collections should be represented, basing his techniques on those developed for meta-search engines. Doulkeridis *et al.* focus more on how to use collection representations to create routing topologies.

In the following industry talk, Ira Woodhead presented an infrastructure for handling massive data sets. The talk by Martin Rajman presented an interesting idea to manage feedback in a distributed way to improve web search, tailoring search results to the user needs.

The seventh and eighth talks considered hierarchical data in query routing (XML based) or general summarization. Venkata Cherukuri presented work on propagation-vectors for trees, a summarization technique to form concise summaries of hierarchical content. This can be used for detecting promising peers for queries. Judith Winter presented an approach to route structured queries in peer-to-peer systems.

Finally, Dongmei Jia presented an automatic approach for spam detection in peer-to-peer file sharing systems.

4 Presentations

Torsten Suel gave a keynote talk on conventional and peer-to-peer search engine architectures [11]. The talk covered the architectural principles of a Web search engine including data acquisition (crawling), data mining, indexing and query processing. In particular, Suel presented several early termination techniques for query processing optimization and also talked about a possibility of leveraging a peer-to-peer architecture for Web search including recent research results and challenges.

Fabrizio Falchi presented the paper on using caching for similarity search [7]. The authors propose the storage of queries and their associated results in a cache with the goal of reusing them to answer future queries. Their idea is to use metric distance functions to determine which of the results in the cache are mathematically guaranteed to be in the K nearest neighbors of a query. When the cache is full, an LRU mechanism is used to flush out queries and their associated results. To control computational complexity, they propose techniques for searching only a subset of the cache.

Experimental results using the metric cache on a synthetic data set derived from photo sharing Web sites are promising. The hit rate increases linearly with the percentage of queries cached. At 5% caching, the hit rate is approximately 20%, with roughly half the hits attributed to each of exact and approximate hits. Result errors - measured by result distance differences from the ideal - are relatively low, being no more than about 10%.

Stanislav Barton presented the paper on self-organized image retrieval [3]. The authors describe the design of their Image Retrieval Network (IRN) system, which allows the content-based search of images from a distributed network of users. The system is based on the dynamic creation of an unstructured network whereas existing works usually address on the distribution of metadata about the shared content in a structured overlay network. The system dynamically creates a network by allowing nodes to create links to other nodes that have returned the most matching results. In this way, a database of the data available on other nodes is created.

Tests of this system were conducted using 10 million annotated images crawled from the Flickr Web site. The authors created a set of 50 queries for the test. These 50 queries were used to initialize the network and then a subset of 20 of these queries was used for experimental measurements. Experimental results indicate that with each successive query, recall increases steadily from 35% to over 75%, while cost, in terms of the percentage of nodes searched per query, is controlled at about 30% of the nodes.

Hans Witschel presented a study on ranking information resources in peer-to-peer text retrieval [13]. The author considers the problem of resource selection in a peer-to-peer environment: given multiple peer profiles, to which subset should a query be routed? Specifically,

the problem of profile creation and ranking is covered. A peer's profile is created by selecting the top-ranked terms from the peer's corpus, ranked using a weighting scheme from the CORI metasearch engine, which counts corpus-dependent and corpus-independent term frequencies. Each neighbor is ranked based on the weights of the query terms in its profile. The query is routed to the top-ranked neighbors. Finally, results are ranked using BM25.

Experiments were conducted on data from CiteSeer, Oshumed and the German GIRT collection. Results indicate that query accuracy using a profile that is pruned to 80 terms is similar to that of an unpruned profile (the sizes of the original profiles are not explicitly given).

Witschel also considers ranking neighbors based on a query expansion technique and profile adaptation based on query results. Query expansion is performed using the local context analysis technique by Xu and Croft [14]. Query expansion proves to reduce accuracy in almost all cases. Profiles are adapted based on how well a particular peer answers a query by increasing the profile weights of the respective query terms. In most cases, the routing accuracy increases by over 10%.

A guest speaker presented the paper on peer-to-peer similarity search [6]. The authors describe a way of creating clusters of peers based on their document collections. Their idea is to represent each peer's local collection by a set of clusters and then to represent each collection by a descriptor (i.e., a document cluster is represented by a term frequency vector). Based on these descriptors peers are clustered forming "Semantic Overlay Networks" (SON). Each SON elects a peer to become the "superpeer" representative, which is in charge of representing the contents of the SON and performing source selection for incoming queries. Query routing is done by comparing a query to a superpeer descriptor for its SON.

Experiments using the GOV2 and Reuters corpora, and the GT-ITM topology generator indicate that the proposed topology can more than triple recall versus a randomly generated super-peer network and increase recall by a factor of more than seven compared with normalized flooding given a fixed cost.

Ira Woodhead presented "Robot Army" [1], a distributed data processing system based on MapReduce [5], supporting data transformation and aggregation operations. In contrast to existing distributed storage solutions, it focuses on I/O using standard input and output streams (STDIN/OUT), and hence, does not come with its own complex application programming interface. Its open source release under the GNU Public License is planned.

Martin Rajman presented recent work on feedback information management for distributed retrieval systems [8]. The approach incorporates user feedback and profiling information in the search process to enhance the query results obtained from users, considering their particular interests. As it has been observed in the past, most peer-to-peer application suffer from the so called bootstrap problem (i.e., the problem of attracting users in an early stage of the system deployment process where the amount of available resources is minor compared to existing centralized solutions). The presented approach can be seen complementary to the existing search solution: the main idea is to manage feedback information in a decentralized fashion and to use this information on a per-user basis during the actual query processing.

Venkata S. Cherukuri presented a tree summarization approach to route requests to suitable nodes in a peer-to-peer network [4]. The approach aims at reducing the amount of metadata needed for a meaningful peer selection to answer user queries. First focusing on the summarization of hierarchical data it is later on explained how to perform similarity queries over these summaries. The talk concluded with a report on a detailed performance evalua-

tion considering multiple data sets, comparing the presented approach to existing works in terms for clustering capabilities.

Judith Winter presented her work on routing of structured queries in peer-to-peer networks [12]. In the talk she addressed the following question: how can structural information help to efficiently route structured queries and thus improve the retrieval of XML documents in a peer-to-peer network? She presented the approach called SPIRIX – Search Engine for P2P Information Retrieval in XML Documents. SPIRIX relies on indexing based on combinations of (content, structure)-tuples called XTerms, resembling indexing with Highly Discriminative Keys (HDKs) [10]. However, in this work not only textual but also structural information is taken into account. As a result, such an index can process "content-only" queries consisting of keywords as well as "content-and-structure" queries. Judith presented preliminary experimental results with INEX collection that show benefits of using structural information.

Dongmei Jia presented work on spam detection in peer-to-peer networks, identifying four different types of spam that is present in nowadays file-sharing systems [9]. The presented approach does not require human attention to the spam detection process by automatically obtaining additional information about files, without downloading the full content. By re-ranking the obtained search results using this additional information, such as the number of local replicas, the user perceived quality of the final ranking is greatly improved. Additional improvements of the approach are presented to further reduce the cost of the basic approach. The experimental evaluation shows that for a typical top-20 query, the amount of spam is reduced by up to 92%.

5 Final Remarks

The presented work is a representative cross section of the research trends in Distributed Information Retrieval. The topics covered include multimedia search [7, 3], distributed query processing [11, 13, 6, 8], high performance distributed processing [1], resource discovery [4, 12] and data quality [9]. The presented solutions show promise in improving the performance and efficiency in current peer-to-peer systems that share user-annotated binary files, multimedia files and text documents. More information including electronic versions of the papers and the presented slides are available at the Workshop Web-site at `http://lsirwww.epfl.ch/LSDS-IR08`.

The next edition of the Large Scale Distributed Systems for Information Retrieval Workshop is planned to be held in conjunction with the 2009 ACM SIGIR Conference in Boston, Massachusetts. There, we plan on continuing the development of new problem areas and solutions for Distributed Information Retrieval. The Workshop Web site can be found at `http://lsdsir09.isti.cnr.it/`.

References

[1] Robotarmy - a distributed system for the casual manipulation of massive data sets. `http://code.google.com/p/robotarmy`.

[2] R. Baeza-Yates, C. Castillo, F. Junqueira, V. Plachouras, and F. Silvestri. Challenges on distributed web retrieval. In *ICDE*, pages 6–20, Istanbul, Turkey, 2007.

[3] S. Barton, V. Dohnal, J. Sedmidubský, and P. Zezula. Building self-organized image retrieval network. In *LSDS-IR*, pages 51–58, Napa, USA, 2008.

[4] V. S. Cherukuri and K. S. Candan. Propagation-vectors for trees (PVT): concise yet effective summaries for hierarchical data and trees. In *LSDS-IR*, pages 3–10, Napa, USA, 2008.

[5] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. In *OSDI*, pages 137–150, San Francisco, USA, 2004.

[6] C. Doulkeridis, K. Nørvåg, and M. Vazirgiannis. Peer-to-Peer similarity search over widely distributed document collections. In *LSDS-IR*, pages 35–42, Napa, USA, 2008.

[7] F. Falchi, C. Lucchese, S. Orlando, R. Perego, and F. Rabitti. A metric cache for similarity search. In *LSDS-IR*, pages 43–50, Napa, USA, 2008.

[8] P. Felber, T. Luu, M. Rajman, and E. Riviere. Managing collaborative feedback information for distributed retrieval. In *LSDS-IR*, pages 27–34, Napa, USA, 2008.

[9] D. Jia. Cost-effective spam detection in P2P file-sharing systems. In *LSDS-IR*, pages 19–26, Napa, USA, 2008.

[10] I. Podnar, M. Rajman, T. Luu, F. Klemm, and K. Aberer. Scalable Peer-to-Peer web retrieval with highly discriminative keys. In *ICDE*, pages 1096–1105, Istambul, Turkey, 2007.

[11] T. Suel. Search engine architectures from conventional to P2P. In *LSDS-IR*, pages 1–2, Napa, USA, 2008.

[12] J. Winter. Routing of structured queries in large-scale distributed systems. In *LSDS-IR*, pages 11–18, Napa, USA, 2008.

[13] H. F. Witschel. Ranking information resources in peer-to-peer text retrieval: an experimental study. In *LSDS-IR*, pages 75–82, Napa, USA, 2008.

[14] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *SIGIR*, pages 4–11, Zurich, Switzerland, 1996.