The 8th Workshop on Large-Scale Distributed Systems for Information Retrieval (LSDS-IR'10)

Roi Blanco
Yahoo! Research
Barcelona, Spain
roi@yahoo-inc.com

B. Barla Cambazoglu
Yahoo! Research
Barcelona, Spain
barla@yahoo-inc.com

Claudio Lucchese
CNR – ISTI
Pisa, Italy
claudio.lucchese@isti.cnr.it

Abstract

The size of the Web as well as user bases of search systems continue to grow exponentially. Consequently, providing subsecond query response times and high query throughput become quite challenging for large-scale information retrieval systems. Distributing different aspects of search (e.g., crawling, indexing, and query processing) is essential to achieve scalability in large-scale information retrieval systems. The 8th Workshop on Large-Scale Distributed Systems for Information Retrieval (LSDS-IR'10) has provided a venue to discuss the current research challenges and identify new directions for distributed information retrieval. The workshop contained two industry talks as well as six research paper presentations. The hot topics in this year's workshop were collection selection architectures, application of MapReduce to information retrieval problems, similarity search, geographically distributed web search, and optimization techniques for search efficiency.

1 Introduction

The Web has been continuously growing. The indexed web is estimated to contain at least 15 billion pages[1]. In the mean time, a dramatic increase is observed in the query traffic of major search engines. Coping with the growth of the Web and query traffic requires scalable information retrieval systems. Today's commercial search engines fully replicate their web index on a few data centers. However, this approach has known scalability problems. It is crucial to come up with scalable indexing and query processing techniques for next generation information retrieval systems of the future. This and similar issues form the focus of the 8th Workshop on Large-Scale Distributed Systems for Information Retrieval (LSDS-IR'10).

This year's workshop (July 23, 2010) was co-located with the 2010 ACM SIGIR Conference in Geneva, Switzerland. The workshop provided an opportunity for the experts in the area of distributed information retrieval to gather and discuss the current research challenges as well as future directions in the field. LSDS-IR'10 continued to build on the efforts of previous workshops:

[1] `http://www.worldwidewebsize.com/`

- LSDS-IR'09: Workshop on Large-Scale Distributed Systems for Information Retrieval (SIGIR'09)
- LSDS-IR'08: Workshop on Large-Scale Distributed Systems for Information Retrieval (CIKM'08)
- LSDS-IR'07: Workshop on Large-Scale Distributed Systems for Information Retrieval (SIGIR'07)
- P2PIR'06: Workshop on Information Retrieval in Peer-to-Peer Networks (CIKM'06)
- P2PIR'05: Workshop on Information Retrieval in Peer-to-Peer Networks (CIKM'05)
- HDIR'05: Workshop on Heterogeneous and Distributed Information Retrieval (SIGIR'05)
- P2PIR'04: Workshop on Information Retrieval in Peer-to-Peer Networks (SIGIR'04)

The rest of the report is organized as follows. We first present the workshop program committee and the program. We then provide an overview of the keynote talks and the research papers presented in the workshop. We conclude the report with some final remarks.

2 Program Committee

The program committee is made up of 19 researchers, who are very active in the field:

- Karl Aberer, Ecole Polytechnique Federale de Lausanne, Switzerland
- I. Sengor Altingovde, Bilkent University, Turkey
- Ricardo Baeza-Yates, Yahoo! Research, Spain
- Ranieri Baraglia, ISTI-CNR, Italy
- Fabrizio Falchi, ISTI-CNR, Italy
- Ophir Frieder, Georgetown University, USA
- Sebastian Michel, Saarland University, Germany
- Kjetil Norvag, Norwegian University of Science and Technology, Norway
- Salvatore Orlando, University of Venice, Italy
- Josiane Parreira, Digital Enterprise Research Institute, Ireland
- Raffaele Perego, ISTI-CNR, Italy
- Vassilis Plachouras, Athens University of Economics, Greece
- Gleb Skobeltsyn, Google, Switzerland
- Torsten Suel, Polytechnic University, USA
- Christos Tryfonopoulos, University of Peloponnese, Greece
- Wai Gen Yee, Illinois Institute of Technology, Chicago USA
- Ivana Zarko, University of Zagreb, Croatia
- Pavel Zezula, Masaryk University of Brno, Czech Republic
- Justin Zobel, NICTA, Australia

3 Workshop Program

The program[2] contained two keynote talks, given by well-known researchers, who are affiliated with leading industry companies. The program also contained six technical paper presentations, one of which is a position paper. The best paper award is given to Tonellotto et al. for their paper, entitled "Efficient dynamic pruning with proximity support". Overall, the talks in the workshop generated quite interesting discussions among the attendees.

3.1 Keynote Talks

- Ricardo Baeza-Yates (Yahoo!), "Towards a distributed web search engine".
- Abdur Chowdury (Twitter), "Twitter".

Towards a distributed web search engine (talk by Baeza-Yates): The scalability of todays commercial search engines is achieved by replicating the web collection and the infrastructure on a number of geographically distant data centers. In addition to benefits in availability, this approach brings benefits in query response times as the distances between users and data centers become shorter. In his talk, Baeza-Yates presented a web search engine architecture that takes this one step further: a geographically distributed multi-site web search engine architecture, where documents are partitioned on data centers based on their relevance to queries issued. He discussed the research problems arising in this architecture, such as query forwarding and data replication. He also provided a number of recent research results, related to web crawling, indexing, query processing, and caching in multi-site search architectures. This talk raised interesting discussions on the public availability of data and software for distributed IR research.

Twitter (talk by Chowdury): Twitter is one of the largest social networking sites in the Internet. It gives millions of people the ability to follow daily activities of other people. In his talk, Chowdury provided an overview of Twitter and demonstrated the scale of the research problems they have been working on. He specifically pointed at the scalability issues in data acquisition and the recency issues caused by the high volume of streaming data.

3.2 Research Paper Presentations

- Almer Tigelaar and Djoerd Hiemstra (University of Twente), "Query-based sampling using snippets".
- Aleksandar Stupar, Sebastian Michel, and Ralf Schenkel (Saarland University), "RankReduce – processing k-nearest neighbor queries on top of MapReduce".
- Anagha Kulkarni and Jamie Callan (Carnegie Mellon University), "Topic-based index partitions for efficient and effective selective search".
- Gianmarco De Francisci Morales (IMT Institute for Advanced Studies), Claudio Lucchese, and Ranieri Baraglia (ISTI-CNR), "Scaling out all pairs similarity search with MapReduce".
- Nicola Tonellotto (ISTI-CNR), Craig Macdonald, and Iadh Ounis (University of Glasgow), "Efficient dynamic pruning with proximity support".

[2]The workshop site is available at `http://www.lsdsir.org/`. The proceedings of the workshop are available at `http://CEUR-WS.org/Vol-630/`.

- Fidel Cacheda, Victor Carneiro, Diego Fernández, and Vreixo Formoso (University of A Coruña), "Performance evaluation of large-scale information retrieval systems scaling down".

Query-based sampling using snippets (talk by Tigelaar) [5]: For effective federated search, it is necessary to accurately estimate the properties of collections indexed in non-cooperative IR systems, which do not voluntarily provide this information to the federator. The collection properties are traditionally estimated by issuing a number of queries to the non-cooperative IR system, retrieving a number of best-matching documents, and then extracting the term distributions in the sampled collection. The main bottleneck in this technique is the high volume of data that needs to be downloaded from the IR system. Tigelaar and Hiemstra presented a novel approach to sample the collection of an IR system, by using the snippets returned in the search results. The presented approach is shown to alleviate the data transfer problem, without significantly sacrificing from accuracy of estimations.

RankReduce – processing k-nearest neighbor queries on top of MapReduce (talk by Stupar) [4]: The scalability becomes an issue when k-nearest neighbor queries are processed over very large datasets. In their work, Stupar et al. considered a scenario, where the dataset is distributed over a cluster of computers. The authors proposed an approach, which combined locality sensitive hashing with the MapReduce framework, to process a batch of k-nearest neighbor queries over the data. The efficiency of the approach is empirically demonstrated on both synthetic and real-life datasets.

Topic-based index partitions for efficient and effective selective search (talk by Callan) [2]: In some web search architectures, a global web index is partitioned into multiple disjoint subindexes. A traditional research problem emerging in such architectures is to reduce the number of subindexes queried without hurting the quality of search results, as much as possible. Kulkarni and Callan presented a comparison of three different index partitioning strategies (random, URL-based, topic-based), coupled with an index selection strategy. Their experiments on three large datasets indicate that the topic-based partitioning strategy reduces the query processing overhead by at least an order of magnitude compared to evaluation over the full index, without any loss in search quality.

Scaling out all pairs similarity search with MapReduce (talk by De Francisci Morales) [3]: Given a collection of objects, the all pairs similarity search problem involves discovering all those pairs of objects whose similarity is above a certain threshold. De Francisci Morales et al. describe a new parallel algorithm that employs the MapReduce framework to scale out to large datasets. The authors explore various pruning techniques that allow reducing the number of evaluated pairs while yielding an exact result. Experiments on real-life data show five-fold improvements in speed over other state-of-the-art approaches.

Efficient dynamic pruning with proximity support (talk by Tonellotto) [6]: Scoring functions used by the web search engines do not rely only on individual term statistics, but also the proximity of query terms in the document. Computing proximity information over the inverted index may bring too much overhead into query processing. Tonellotto et al. propose an early termination technique for document-at-a-time ranking models that use proximity information. Experiments over a large document collection have demonstrated significant performance benefits, especially for long queries.

Performance evaluation of large-scale information retrieval systems scaling down (talk by Cacheda) [1]: Evaluating the efficiency of a search engine is typically done in one of the three ways: analytical models, simulations, or empirical studies using a real search engine. Cacheda et al. illustrated the difficulties involved in these three potential options. The

authors proposed a technique based on virtualization to create a scaled-down search engine to imitate a real one. The main idea in this technique is to maintain the overall performance behavior of a real search engine while reducing the computational requirements.

4 Final Remarks

This year's workshop papers encompassed a variety of topics; some of the presented works are follow-up on well-established research issues, ranging from efficient query processing [6] to resource selection on federated environments [2] and collection statistics estimation (using snippet sampling) [5]. On perspective, map-reduce applications are gaining an increasing interest in the community, especially those that involve similarity computations [3, 4]. The map-reduce programming paradigm allows experimentation and development of novel techniques for old problems in a new framework. In general, there is opportunity for exploring topics and solutions that were complicated to scale beforehand, and therefore making large-scale distributed computation more accessible, whilst facing new interesting problems. Finally, it is clear that there is a need in academia to keep up with real data corpora sizes and to experiment with the high number of machines employed in real-world commercial scenarios, which is not feasible in most of the situations [1].

5 Sponsors

We are grateful to Yahoo! and Twitter for their sponsorship.

References

[1] F. Cacheda, V. Carneiro, D. Fernández, and V. Formoso. Performance evaluation of large-scale information retrieval systems scaling down. In *Proceedings of the 8th Workshop on Large-Scale Distributed Systems for Information Retrieval*, pages 36–39. CEUR WS, 2010.

[2] A. Kulkarni and J. Callan. Topic-based index partitions for efficient and effective selective search. In *Proceedings of the 8th Workshop on Large-Scale Distributed Systems for Information Retrieval*, pages 19–24. CEUR WS, 2010.

[3] G. D. F. Morales, C. Lucchese, and R. Baraglia. Scaling out all pairs similarity search with MapReduce. In *Proceedings of the 8th Workshop on Large-Scale Distributed Systems for Information Retrieval*, pages 25–30. CEUR WS, 2010.

[4] A. Stupar, S. Michel, and R. Schenkel. RankReduce – processing k-nearest neighbor queries on top of MapReduce. In *Proceedings of the 8th Workshop on Large-Scale Distributed Systems for Information Retrieval*, pages 13–18. CEUR WS, 2010.

[5] A. Tigelaar and D. Hiemstra. Query-based sampling using snippets. In *Proceedings of the 8th Workshop on Large-Scale Distributed Systems for Information Retrieval*, pages 7–12. CEUR WS, 2010.

[6] N. Tonellotto, C. Macdonald, and I. Ounis. Efficient dynamic pruning with proximity support. In *Proceedings of the 8th Workshop on Large-Scale Distributed Systems for Information Retrieval*, pages 31–35. CEUR WS, 2010.