Scalability and Efficiency Challenges in Large-Scale Web Search Engines

Ricardo Baeza-Yates
Yahoo Labs
Barcelona, Spain
rbaeza@acm.org

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

ABSTRACT

The main goals of a web search engine are quality, efficiency, and scalability. In this tutorial, we focus on the last two goals, providing a fairly comprehensive overview of the scalability and efficiency challenges in large-scale web search engines. In particular, the tutorial provides an in-depth architectural overview of a web search engine, mainly focusing on the web crawling, indexing, and query processing components. The scalability and efficiency issues encountered in these components are presented at four different granularities: at the level of a single computer, a cluster of computers, a single data center, and a multi-center search engine. The tutorial also points at open research problems and provides recommendations to researchers who are new to the field.

Categories and Subject Descriptors

H.3.3 [**Information Storage Systems**]: Information Retrieval Systems.

Keywords

Web search engines; crawling; indexing; query processing; caching; efficiency; scalability.

1. DESCRIPTION

The content of this tutorial is mostly based on a book chapter published by the presenters [2]. A subset of the tutorial was previously presented in SIGIR'13 [3]. The main parts of this updated version are summarized next.

1.1 Main Concepts

First, we explain the main challenges posed by the Web and its characteristics. Second, we present the main components of a search engine and possible software/hardware architectures. Third, we summarize the interactions among the three main components: crawler, indexer, and query processor. Finally, we cover caching, replication, and index partitioning as crucial techniques behind efficiency and scalability.

1.2 Web Crawling

A web crawler is responsible for discovering new web pages and downloading their content while refreshing the content of previously downloaded pages [4]. The main performance objective for a crawler is to attain a high page download rate. The efficiency of a sequential web crawler is mainly determined by the selection and implementation of proper data structures. Scalability is commonly achieved through multi-threading and parallelization. In addition to such standard techniques, our tutorial covers geographically distributed web crawling, which emerges as a promising option for further scalability, and discusses the web partitioning and crawler placement problems.

1.3 Indexing

The indexing process involves various parsing, extraction, and classification tasks, through which certain features are extracted and the textual content of downloaded pages are converted into an inverted index [5]. The main performance metrics for an indexer are the length of deployment cycles, compactness, and speed of index updates. So far, most algorithmic improvements are concentrated on index compression. At the architectural level, the efficiency and scalability are tried to be improved via index partitioning, pruning, and replication. Our tutorial also covers the indexing strategies for geographically distributed search engines.

1.4 Query processing

Query processing aims to generate the best-matching result set for a given query [1]. The efficiency of a query processor is assessed by its throughput and average response latency. In addition to the state-of-the-art ranking techniques employed in web search engines, our tutorial covers a variety of problems including query processing on multi-core architectures, early exit optimizations, skipping in inverted lists, and query forwarding. The tutorial also covers common issues in caching, such as admission, eviction, and prefetching while focusing on the cache freshness problem, which has recently attracted attention.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577271.

2. OBJECTIVES

The following are the main objectives of the tutorial.

- To provide an in-depth background on the architectural components of a web search engine.
- To present the fundamental scalability and efficiency issues which have been often addressed in the information retrieval literature.
- To shed some light into the techniques used in large-scale commercial search engines and bridge the gap between the industry and academia.
- To identify open research problems in the context of web search engine scalability and efficiency, promoting further research on the topic.

3. PRESENTERS

Ricardo Baeza-Yates is VP of Yahoo Labs for Europe and Latin America, leading the labs at Barcelona, Spain and Santiago, Chile. Until 2005, he was the director of the Center for Web Research at the Department of Computer Science of the Engineering School of the University of Chile, and ICREA Professor at the Dept. of Information and Communication Technologies of University Pompeu Fabra in Barcelona, Spain. He is co-author of the bestseller textbook Modern Information Retrieval by Addison-Wesley, first published in 1999 with a second edition in 2011, as well as co-author of the second edition of the Handbook of Algorithms and Data Structures, Addison-Wesley, 1991; and co-editor of Information Retrieval: Algorithms and Data Structures, Prentice-Hall, 1992, among more than 400 other publications. He has been PC-Chair of the most important conferences in the field of Web Search and Web Mining. He has given tutorials in most major conferences many times, including SIGIR, WWW and VLDB. He has won several awards and is, both, ACM and IEEE Fellow.

Berkant Barla Cambazoglu received his BS, MS, and PhD degrees, all in computer engineering, from the Computer Engineering Department of Bilkent University in 1997, 2000, and 2006, respectively. He has then worked as a post-doctoral researcher in the Biomedical Informatics Department of the Ohio State University. He is currently employed as a senior researcher in Yahoo Labs, where he is heading the web retrieval group. He has many papers published in prestigious journals including IEEE TPDS, JPDC, Inf. Syst., ACM TWEB, and IP&M, as well as top-tier conferences, such as SIGIR, CIKM, WSDM, WWW, and KDD.

4. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Publishing Company, USA, 2nd edition, 2011.

[2] B. B. Cambazoglu and R. Baeza-Yates. Scalability challenges in web search engines. In M. Melucci, R. Baeza-Yates, and W. B. Croft, editors, *Advanced Topics in Information Retrieval*, volume 33 of *The Information Retrieval Series*, pages 27–50. Springer Berlin Heidelberg, 2011.

[3] B. B. Cambazoglu and R. Baeza-Yates. Scalability and efficiency challenges in commercial web search engines. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1124, 2013.

[4] C. Olston and M. Najork. Web crawling. *Foundations and Trends in Information Retrieval*, 4(3):175–246, 2010.

[5] J. Zobel and A. Moffat. Inverted files for text search engines. *ACM Computing Surveys*, 38(2), 2006.