Invited Abstract

Towards a Distributed Search Engine

Ricardo Baeza-Yates
Yahoo! Labs
Barcelona, Spain
Email: http://www.baeza.cl

Abstract

In the dynamic ocean of web data, where we have over 200 million websites, web search engines are the primary way to access content. As the data is on the order of petabytes, current search engines are very large centralized systems based on replicated clusters, where easily more than 100 billion web pages are indexed. On the other hand, Internet users are above two billion and hundreds of million of queries are issued each day. In the near future, centralized systems are likely to become less effective against such a data-query load, thus suggesting the need of fully distributed search engines. Such engines need to maintain high quality answers, fast response time, high query throughput, high availability and scalability; in spite of network latency and scattered data. In this talk we present the main challenges behind the design of a distributed web retrieval system and our research in all the components of a search engine: crawling, indexing, and query processing.

SUMMARY

The main challenge is to mimic a centralized web search engine [5, Chapter 11 with Yoelle Maarek] with a distributed search engine [9] in spite of the network latency. Hence, our baseline is to obtain the same results of the centralized architecture. In addition, to make the distributed architecture interesting, we require that the new web search engine must be less expensive than the centralized one. This motivated the development of a cost model for web search engines [4]. We show that the above goal is feasible and that the key components in the solution are smart caching algorithms [1], [2], [10], prediction techniques based in machine learning [1], [3], replicating the most used data as close as possible to the users needing it [4], [6], [12], and using new distributed query processing techniques [7], [8], [11].

REFERENCES

[1] R. Baeza-Yates, F. Junqueira, V. Plachouras and H.F. Witschel. Admission Policies for Caches of Search Engine Results, In SPIRE 2007, LNCS 4726, Springer, Santiago, Chile, 74–85, 2007.
[2] R. Baeza-Yates, A. Gionis, F. Junqueira, V. Murdock, V. Plachouras, and F. Silvestri. The Impact of Caching on Search Engines. In 30th Annual International ACM SIGIR Conference, Amsterdam, Netherlands, 183–190, 2007
[3] R. Baeza-Yates, V. Murdock, and C. Hauff. Efficiency Trade-offs in Twotier Web Search Systems. In SIGIR 2009, Boston, USA, 163–170, 2009.
[4] R. Baeza-Yates, A. Gionis, F. Junqueira, V. Plachouras, and L. Telloli. On the Feasibility of Multi-Site Web Search Engines (best paper award). In ACM CIKM 2009, Hong Kong, China, 425–434, 2009.
[5] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, 2ed ed. Harlow, England: Addison-Wesley, 2011.
[6] R. Blanco, B.B. Cambazoglu, F. Junqueira, I. Kelly, and V. Leroy. Assigning documents to master sites in distributed search. In CIKM 2011, Glasgow, UK, 67–76, 2011.

[7] B.B. Cambazoglu, V. Plachouras, and R. Baeza-Yates. Quantifying Performance and Quality Gains in Distributed Web Search Engines. In SIGIR 2009, Boston, USA, 411–418, 2009.
[8] B.B. Cambazoglu, E. Varol, E. Kayaaslan, C. Aykanat, and R. Baeza- Yates. Query forwarding in geographically distributed search engines. In SIGIR 2010, Geneve, Switzerland, 90–97, 2010.
[9] B.B. Cambazoglu, and R. Baeza-Yates. Scalability Challenges in Web Search Engines. In Advanced Topics in Information Retrieval, M. Melucci and R. Baeza-Yates (eds), Springer, 27-50, 2011.
[10] B.B. Cambazoglu, I.S. Alting¨ovde. Impact of Regionalization on Performance of Web Search Engine Result Caches. In SPIRE 2012, Cartagena de Indias, Colombia, 161–166, 2012.
[11] E. Kayaaslan, B.B. Cambazoglu, R. Blanco, F. Junqueira, and C. Aykanat. Energy-price-driven query processing in multi-center web search engines. In SIGIR 2011, Beijing, China, 983–992, 2011.
[12] E. Kayaaslan, B.B. Cambazoglu, and C. Aykanat. Document replication strategies for geographically distributed web search engines. Information Processing & Management 49(1): 51–66, 2013.