Towards a Distributed Web Search Engine

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

Abstract: We present recent and on-going research towards the design of a distributed Web search engine. The main goal is to be able to mimic a centralized search engine with similar quality of results and performance, but using less computational resources. The main problem is the network latency when different servers have to process the queries. Our preliminary findings mix several techniques, such as caching, locality prediction and distributed query processing, that try to maximize the fraction of queries that can be solved locally.

1 Summary

Designing a distributed Web search engine is a challenging problem [BYCJ+07], because there are many external factors that affect the different tasks of a search engine: crawling, indexing and query processing. On the other hand, local crawling profits with the proximity to Web servers, potentially increasing the Web coverage and freshness [CPJT08]. Local content can be indexed locally, communicating later local statistics that can be helpful at the global level. So the natural distributed index is a document partitioned index [BYRN99].

Query processing is very efficient for queries that can be answered locally, but too slow if we need to request answers from remote servers. One way to improve the performance is to increase the fraction of queries that look like local queries. This can be achieved by caching results [BYGJ+08a], caching partial indexes [SJPBY08] and caching documents [BYGJ+08b], with different degree of effectiveness. A complementary technique is to predict if a query will need remote results and request in parallel local and remote results, instead of doing a sequential process [BYMH08]. Putting all these ideas together we can have a distributed search engine that has similar performance to a centralized search engine but that needs less computational resources and maintenance cost than the equivalent centralized Web search engine [BYGJ+08b].

Future research must study how all these techniques can be integrated and optimized, as we have learned that the optimal solution changes depending on the interaction of the different subsystems. For example, caching the index will have a different behavior if we are caching results or not.

References

[BYCJ+07] Ricardo Baeza-Yates, Carlos Castillo, Flavio Junqueira, Vassilis Plachouras and Fabrizio Silvestri. Challenges on Distributed Web Retrieval. In *ICDE*, 6–20, 2007.

[BYGJ+08a] Ricardo Baeza-Yates, Aristides Gionis, Flavio P. Junqueira, Vanessa Murdock, Vassilis Plachouras and Fabrizio Silvestri. Design trade-offs for search engine caching. *ACM Trans. Web*, 2(4):1–28, 2008.

[BYGJ+08b] Ricardo Baeza-Yates, Aristides Gionis, Flavio P. Junqueira, Vassilis Plachouras and Luca Telloli. On the feasibility of multi-site Web search engines. Submitted, 2008.

[BYMH08] Ricardo Baeza-Yates, Vanessa Murdock and Claudia Hauff. Speeding-Up Two-Tier Web Search Systems. Submitted, 2008.

[BYRN99] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, May 1999.

[CPJT08] B. Barla Cambazoglu, Vassilis Plachouras, Flavio Junqueira and Luca Telloli. On the feasibility of geographically distributed web crawling. In *InfoScale '08: Proceedings of the 3rd international conference on Scalable information systems*, 1–10, ICST, Brussels, Belgium, Belgium, 2008.

[SJPBY08] Gleb Skobeltsyn, Flavio Junqueira, Vassilis Plachouras and Ricardo Baeza-Yates. ResIn: a combination of results caching and index pruning for high-performance web search engines. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 131–138, New York, NY, USA, 2008. ACM.