Towards a Distributed Web Search Engine

Ricardo Baeza-Yates
Yahoo! Research Barcelona Espagne
ricardo.baeza@barcelonamedia.org
http://www.dcc.uchile.cl/∼rbaeza/

Summary

In the ocean of Web data, 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. Web data, however, is always evolving. The number of Web sites continues to grow rapidly (230 millions at the end of 2009) and there are currently more than 20 billion indexed pages. On the other hand, Internet users are above one 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.

Bibliography

Ricardo Baeza-Yates is VP of Yahoo ! Research for Europe, Middle East and Latin America, leading the labs at Barcelona, Spain and Santiago, Chile, as well as supervising the newer lab in Haifa, Israel. 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 Technology of the Univ. Pompeu Fabra in Barcelona, Spain. He is co-author of the best-seller book Modern Information Retrieval, published in 1999 by Addison-Wesley with a second edition coming in 2010, as well as co-author of the 2nd 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 200 other publications. He has received the Organization of American States award for young researchers in exact sciences (1993) and with two Brazilian colleagues obtained the COMPAQ prize for the best CS Brazilian research article (1997). In 2003 he was the first computer scientist to be elected to the Chilean Academy of Sciences. During 2007 he was awarded the Graham Medal for innovation in computing, given by the University of Waterloo to distinguished ex-alumni. In 2009 he was awarded the Latin American distinction for contributions to CS in the region and became an ACM Fellow.