Distributed Web Retrieval

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

ABSTRACT

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 (over 270 millions at the beginning of 2011) 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 tutorial we present the architecture of current search engines and we explore the main challenges behind the design of all the processes of a distributed Web retrieval system crawling, indexing, and query processing.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search Process

General Terms

Algorithms, Management, Performance

Keywords

Web Retrieval, Distributed Systems, Crawling, Indexing, Query Processing

Contents

In this tutorial we cover the following topics:

- Web search: concepts, challenges.
- Web search architectures: centralized, replicated, distributed.
- Crawling: centralized, distributed.
- Caching: results, index, documents.
- Index partitioning: document based, term based.
- Query processing: collection selection, collection prediction, routing queries.

The content is mainly based in the corresponding chapters of [1] and the research of the presenter.

REFERENCE

[1] Ricardo Baeza-Yates, Berthier Ribeiro-Neto. *Modern Information Retrieval: The Concepts and Technology behind Search Engines*, second edition, Addison-Wesley, 2010.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.