

Turning Data into Knowledge: Challenges and Opportunities at Baidu.com

William I. Chang

Baidu.com

Beijing 100080, P.R.C.

86-10-82606900

wchang@baidu.com

ABSTRACT

The Chinese language search engine Baidu.com is the fourth most trafficked web site in the world, in what will likely soon become the world's largest user-base, China. We will outline three key systems: search engine, iknow.baidu.com, and advertising platform, and describe open problems encountered in the continuing process to improve Baidu.com's technology and services.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *clustering, information filtering, query formulation, retrieval models, search process.*

General Terms

Algorithms, Economics, Management, Measurement, Human Factors, Performance, Reliability.

Keywords

Search engine, question-answering, online advertising, VLDB, WWW, NLP, Asian languages, China.

1. SUMMARY

The Chinese language search engine Baidu.com is the fourth most trafficked web site in the world, in what will likely soon become the world's largest user-base, China. The most impressive

technical feat at Baidu.com is not only a multi-terabyte search engine database, but the fact that the database is refreshed each week. Compared against other search engines, the timeliness of the Baidu.com database is immediately apparent to even casual users. An assembly-line process retrieves content from the WWW, parses and filters the content based on quality assessments, creates indices based on best-of-class NLP techniques, then combines the indices into both physical and logical clusters for deployment on a large network of thousands of servers located across several data centers.

A little over a year ago, Baidu.com launched a question-and-answer community service called iknow.baidu.com. It has logged a monthly average of over one million questions successfully answered, and the database of answers has become an authoritative resource with explosive growth. Q&A and keyword search are complementary in terms of user behavior and are enjoying increasing degrees of traffic cross-flow and overall synergy.

Baidu.com's search and network advertising business models are important drivers of China's economic growth, with over 100,000 customers. The underlying technology is a massive data management system providing real-time ad selection as well as timely, broad, and deep analysis of performance and behavior.

This talk will include open problems encountered in the continuing process to improve Baidu.com's search engine, iknow.baidu.com, and advertising technology.