The Information Discovery Graph: A Framework for a Distributed Search Engine

Nelson Tang and Lixia Zhang
UCLA Department of Computer Science
{ tang, lixia }@cs.ucla.edu

The World Wide Web is an enormous collection of information, but to fully exploit its power, users must be able to find the information they want from that space. Without a doubt, the de facto standard for information discovery on the Web is the search engine. Given the amazing rate of growth of the amount of information available on the Web, search engines are becoming an essential part of the network infrastructure. Therefore it is critical to keep them well-maintained and robust against possible failures.

Currently, nearly all major search engines follow a model where their workloads are distributed among a cluster of workstations; however, the cluster itself is located in a single network location [1]. This exposes that location as being a possible single point of failure. Another potentially very serious problem, though not frequently addressed technically, is the amount of control that is inherent in a single corporation or entity owning a search engine. Given that search engines are becoming required gateways to access the information on the Web, such control by centralized authorities can lead to editorial decisions that affect what content is accessible to the public [3] [2].

One solution to these problems is to avoid the centralized engine model, and one such alternative is a distributed, decentralized model. Using a distributed model, there are no single points of failure or single bottleneck locations; the system is spread out across the entire network. Additionally, with no central authority controlling the search engine, no single entity can make unilateral editorial decisions that affect the entire search space.

We explore a distributed model with the Information Discovery Graph (IDG). The IDG is a framework for a distributed search engine system. The idea is to partition the total search space using a semantic taxonomy into portions which are distributed to different servers. Each server is a *manager*, and it is assigned a topic and is in charge of that piece of the total information space. *Data sources* represent providers of information, such as Web sites; they find the manager responsible for their topics and register themselves with that manager. A user is represented by a *client*, and the queries it sends to the system are transparently routed to the most topically-relevant manager.

Our IDG design focuses on the issues of scalability, adaptability and stability of the distributed system. The challenge comes from maintaining the system's performance as the load on the system increases; this load is made up of *registration load* (the number of data sources registering with the system, i.e., the growth of the search space) and *query load* (the number of clients issuing queries to the system). We have developed a simulation of the IDG system and used actual Web search engine traces to drive the simulation. The results from our investigations so far have validated our design for its scalability, in terms of limiting the per-server overhead to a low, near-constant value, and its performance, in providing logarithmic search times as the registration load increases.

We are currently in the process of continuing to test our design for other measurements, and we hope to demonstrate the viability of the IDG and the decentralized, distributed search engine model.

References

[1] U. Hölzle, *The Google Linux Cluster*. Talk given at Univ. of Washington at Seattle, Nov. 5, 2002. http://www.cs.washington.edu/info/videos/asx/colloq/UHoelzle_2002_11_05.asx

[2] J. Hu, "Yahoo Yields to Chinese Web Laws", *CNET News.com*, Aug. 13, 2002. http://news.com.com/2100-1023-949643.html?tag=rn

[3] A. Orlowski, "Google News: Press Releases are OK - Official", *The Register* (UK), Apr. 5, 2003. http://www.theregister.co.uk/content/6/30112.html