The Information Discovery Graph

A Distributed Search Engine Framework

Goal: build a decentralized, distributed search engine framework
- no single point of failure
- not controlled by any one administration or corporation

Design challenge:
- scalable, robust, and adaptable to changing topic popularity

Approach:
- partition the search space by semantic topic using a hierarchical taxonomy
- generic topics are higher up, more specific topics are lower

Components of the IDG System:
- manager - assigned a topic from the taxonomy, holds listings of data sources with that topic
- topical group - groups together related managers
- data source - represents an information provider (e.g., a website)
- client - searches for managers that have listings of interesting data sources
- cache - helps clients find popular managers quickly

How a query is answered:
- client issues query to its local cache
- cache searches internal memory, sending query to best matching IDG manager
- IDG finds true best manager and responds to cache, who relays response to client

Maintaining the IDG directory:
- managers periodically multicast heartbeat messages
- managers learn about other managers for failure recovery and to help forward queries

Associate topics with locations to reduce heartbeat bandwidth:
- Los Angeles
- New York
- managers use locally-scoped multicast to limit their heartbeats
- faraway managers use a proxy to maintain a presence in the other scope
- a unicast channel connects the manager and proxy

IDG directory adapts to changing topic popularity:
- science manager is overloaded with too many data sources
- new manager from pool of free managers is activated and assigned topic physics
- data sources moved to new manager

Simulation configuration:
- implemented using Parsec language
- Experiments search engine trace over 24 hours; approx. 2.5 million queries, 537,000 unique users (IP addresses)
- queries hashed into a manually-built taxonomy based on Yahoo directory
- to simulate data source registration, queries treated as data sources

Future work:
- measure effects of enhancements: system-wide "Hot Topics" cache, cross-references, duplicate query detection
- other trace data: UCLA traffic, more traces needed!

Summary:
- IDG is framework for decentralized, distributed search engine
- semantic taxonomy provides intuitive browsing
- design addresses scalability, adaptability, and robustness

Nelson Tang (tang@cs.ucla.edu) and Lixia Zhang (lixia@cs.ucla.edu)
http://irl.cs.ucla.edu/IDG/