The Information Discovery Graph

A Distributed Search Engine Framework

INTERNET RESEARCH LAB

Goal: build a decentralized, distributed search engine framework

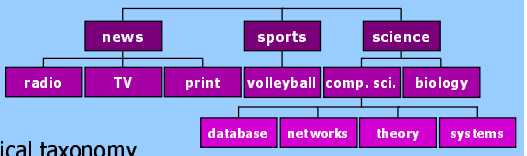
- no single point of failure
- not controlled by any one administration or corporation

Design challenge:

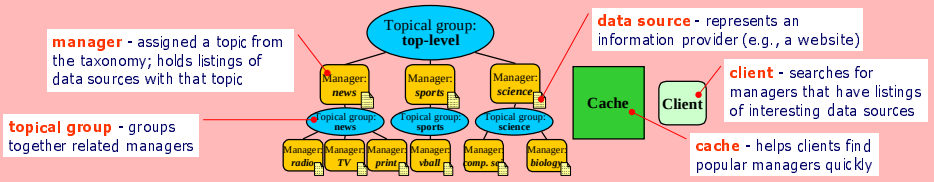
- scalable, robust, and adaptable to changing topic popularity

Approach:

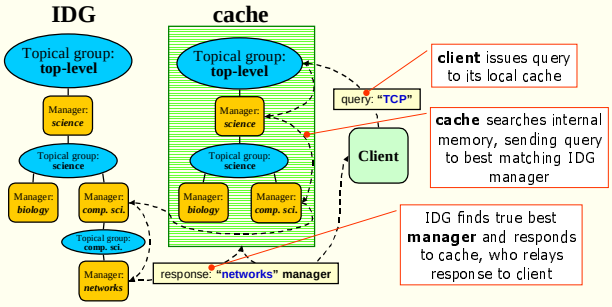
- partition the search space by **semantic topic** using a hierarchical taxonomy
  - generic topics are higher up, more specific topics are lower



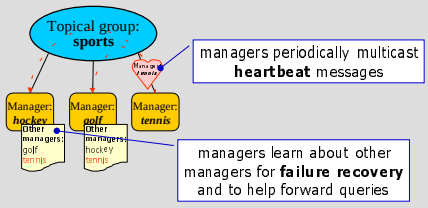
Components of the IDG System:



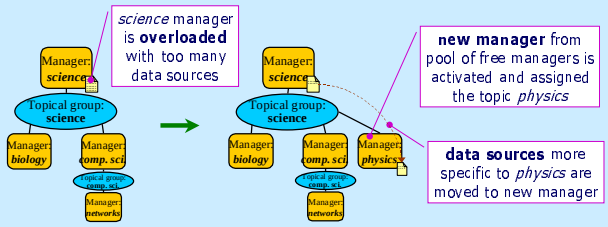
How a query is answered:



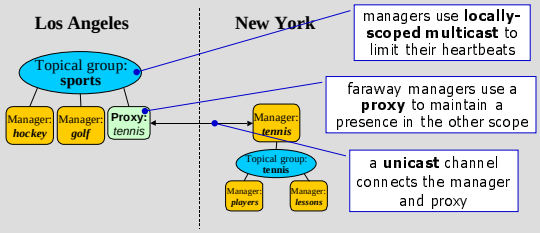
Maintaining the IDG directory:



IDG directory adapts to changing topic popularity:

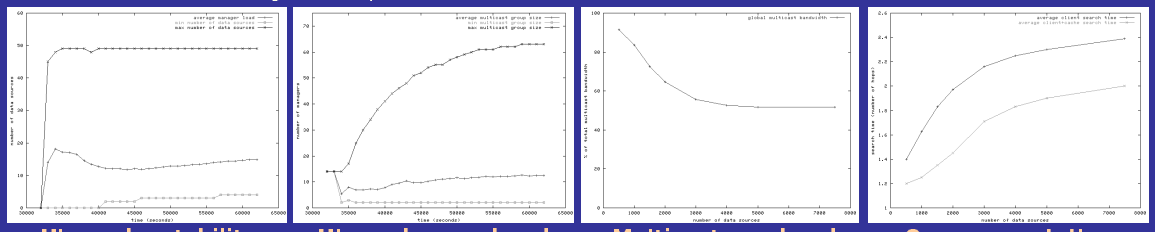


Associate topics with locations to reduce heartbeat bandwidth:



Simulation configuration:

- implemented using **Parsec** language
- Excite 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



Hierarchy stability	Hierarchy overhead	Multicast overhead	Query search time
# of data sources per manager	# of managers per group	% of total multicast with global scope	# of hops per query

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 dencentralized, 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/