Keyword search is the "needle," 20 billion web pages is the "haystack" and advanced search is the "pitch fork."

Context-Based Searching

monograph | solution | summer 2006

CHALLENGE
The average query length in a leading search engine is 2.21 words per search. 7.8% of queries use Boolean operators and only 35% of those searches were correctly issued. (Eastman & Jansen, 2003).

STRATEGY
Upload 40, 60 or even 100 descriptive tokens creating a subset (context) before user enters 2.21 average search terms.

STRATEGY
Context search relies on "thin-slicing" — not limited to building a vocabulary for each course, but detailed categorization algorithms that group articles into multiple buckets to drive "related-retrieval."

STRATEGY
A key to building context sensitive search is a robust "alias" engine. Searching on M.I.T. and MIT often produces very different results, even though the meaning is the same.

The future of search is to create subsets (slices) of a data store based on the context of who is conducting the search. "Know your customer" is a phrase long overdue in search and will revolutionize knowledge discovery.

BASIC SEARCH	*ROBUST SEARCH*	*NON-OBVIOUS SEARCH*	*MANAGED SEARCH*	*SEQUENCED RESULTS*	*NEURAL SEARCH*
media and technology	media and technology and philadelphia and city and neighborhood	[add] urban and planning and knowledge and development and design and community	subtract entity (philadelphia) weight words (media + technology) add taxonomy (M.I.T. course)	pattern matching proximity frequency prominence	weight users view patterns within a thin slice or college course

note: example course can be found at: http://ocw.mit.edu/OcwWeb/Urban-Studies-and-Planning/11-310JSpring2002/Syllabus/index.htm

I am taking a course at M.I.T. on media and technology and decided to do some research for an upcoming assignment. Searching for [media and technology] in Google returns 1.5 billion results. Searching a leading online library returns 48,000 results. The top result in Google takes me to the Jones Encyclopedia, which errors out. The library search takes me to a document on voting.

There have been attempts to address this — see Google Scholar (1.2 million search results) and CiteSeer (4,100 results). But the solution is to not build more search tools. It's to change the way we approach search from keyword to context (+ keyword).

What if I could upload a syllabus (assignment) for the class into a search engine before entering a single keyword? Parsing the syllabus would reveal my class is predominantly about cities, neighborhoods and more specifically, Philadelphia. Just adding these three keywords to the above search reduces the results to 9,000, down from 48,000. Next, let's take into consideration my class is at M.I.T. and weight articles published by M.I.T. professors.

Imagine M.I.T. extending OpenCourseWare to include relevant library documents based on syllabus content for 1,400+ courses — resulting in a recommended reading list and custom "library" for each class.

Everyone wins. Students begin with a smaller footprint for conducting research. Libraries reach more students in a meaningful way.

Schools raise expectations for performance. And someday, professors will use the same system to design and "test" course concepts before releasing to the students.

The value does not stop there. For example, the best way to test and validate a thesis subject is to look for patterns among millions of published articles. And the same system can be used to identify plagiarism by comparing key phrases from one document to the next. Universities will begin to pick up the cost of these systems and institutionalize their use across the campus.

Probably the greatest benefit — once students graduate, they will continue to leverage the library research system (LRS) in their professional careers.

INFORNAUTICS