

Fine-Grained Metasearching of the World Wide Web



James C. French, Allison L. Powell, Edward K. O'Neil
 University of Virginia Department of Computer Science
 Information Retrieval Group
<http://www.cs.virginia.edu/~cyberia>



Conventional metasearching of the WWW is coarse-grained. Queries are sent to several search engines, processed against all the content indexed by the search engines, and finally the separate result lists are merged for presentation to the user.

Fine-grained metasearching is an attempt to improve retrieval effectiveness by targeting queries to subsets of content rather than to an entire collection. For example, we are testing this technique by treating the subject categories available at search engines as separate, individually searchable information resources.

Category Structure of Search Engines

AltaVista	Infoseek	NorthernLight	Yahoo!	Lycos	Snap
arts	automotive	arts	arts	arts	arts
auto	business finance	auto	auto	auto	business
business	careers	careers	careers	careers	careers
computers	computing	contemporary/life	contemporary/life	contemporary/life	computing
education	education	education	education	education	education
entertainment	entertainment	entertainment	entertainment	entertainment	entertainment
family	family	family	family	family	family
games	games	games	games	games	games
health fitness	health	health	health	health	health
home family	home family	home family	home family	home family	home family
internet	internet	internet	internet	internet	internet
living	living	living	living	living	lifestyle
marketplace	marketplace	marketplace	marketplace	marketplace	lifestyle
money	money	money	money	money	lifestyle
news media	news	news	news	news	multimedia
news media	news	news	news	news	news
politics	politics	politics	politics	politics	politics
reference	reference	reference	reference	reference	reference
regional	regional	regional	regional	regional	regional
science	science	science	science	science	retailers
services	services	services	services	services	retailers
sports	sports	sports	sports	sports	sports
society and culture	society and culture	society and culture	society and culture	society and culture	sports
travel	travel	travel	travel	travel	technology
world	world	world	world	world	travel

Query: "lyme disease"

Query: "jupiter moons"

Selection Broker

Our approach involves a *selection broker* that is used to determine what information resources are most useful for a query. The selection broker has a representation of each candidate information resource. The example depicted here shows two queries along with the top 10 ranked information resources for each. The rank of each information resource is shown in parentheses.

Note that the "jupiter moons" query (shown in blue) has been restricted to a subset of the available information resources. In the example here, the last two columns of information resources have been eliminated. This can occur for a variety of reasons, for example, a user may not have access to some information resources (e.g. may not be a subscriber), or may simply elect to have some excluded because of perceived poor quality.

As in conventional metasearching, the result lists returned from each information resource are merged and returned to the user so that the actual web pages can be viewed.

AltaVista	Infoseek	NorthernLight	Yahoo!	Lycos	Snap	AltaVista	Infoseek	NorthernLight	Yahoo!	Lycos	Snap
arts (7)	automotive	arts	arts	arts	arts	arts (7)	automotive	arts	arts	arts	arts
auto	business finance	auto	auto	auto	business	auto	business finance	auto	auto	business	business
business	careers	careers	careers	careers	careers	business	careers	careers	careers	careers	careers
computers	computing	contemporary/life	contemporary/life	contemporary/life	computing	computers	computing (10)	contemporary/life	contemporary/life	contemporary/life	computing
education	education	education	education	education	education	education (9)	education	education	education	education	education
entertainment	entertainment	entertainment	entertainment	entertainment	entertainment	entertainment (9)	entertainment	entertainment	entertainment	entertainment	entertainment
family	family	family	family	family	family	family	family	family	family	family	family
games	games	games	games	games	games	games	games	games	games	games	games
health fitness	health	health	health	health	health	health fitness (8)	health	health	health	health	health
home family	home family	home family	home family	home family	home family	home family	home family	home family	home family	home family	home family
internet	internet	internet	internet	internet	internet	internet	internet	internet	internet	internet	internet
living	living	living	living	living	lifestyle	living	living	living	living	living	lifestyle
marketplace	marketplace	marketplace	marketplace	marketplace	lifestyle	marketplace	marketplace	marketplace	marketplace	marketplace	lifestyle
money	money	money	money	money	lifestyle	money	money	money	money	money	lifestyle
news media	news	news	news	news	multimedia	news media (1)	news	news	news	news	multimedia
news media	news	news	news	news	news	news media (1)	news	news	news	news	news
politics	politics	politics	politics	politics	politics	politics	politics	politics	politics	politics	politics
reference	reference	reference	reference	reference	reference	reference	reference	reference	reference	reference	reference
regional	regional	regional	regional	regional	regional	regional	regional	regional	regional	regional	regional
science	science	science	science	science	retailers	science (1)	science	science	science	science	retailers
services	services	services	services	services	retailers	services	services	services	services	services	retailers
sports	sports	sports	sports	sports	sports	sports	sports	sports	sports	sports	sports
society and culture	society and culture	society and culture	society and culture	society and culture	sports	society and culture	society and culture	society and culture	society and culture	society and culture	sports
travel	travel	travel	travel	travel	technology	travel	travel	travel	travel	travel	technology
world	world	world	world	world	travel	world	world	world	world	world	travel

Information Resources

