Comparison of Metasearch Engines

Selection

I have chosen to compare the following three metasearch engines:

1. QuadSearch - the source of some of the more creative ranking algorithms of the literature I reviewed.

2. Dogpile - one of the more popular metasearch engines.

3. mamma - another metasearch engine, owned by Copernic Inc., a company which specializes in Search-based software.

I have had a bit of experience programming in an Open-Source Statistical scripting language named "R." I've empirically had a very difficult time being able to search for help on this program, because "R" is such an awkwardly-indexed search term. I've been able to find a domain-specific site which I generally use, but I'm curious if the advertised benefit of "clustering" will be able to overcome this problem for me.

I've entered the search term "R global variable" (without quotes) into each metasearch engine and will analyze (primarily) the top 10 results of each.

Initial Perceptions:

I initially scanned the results to see which are specific to my actual interests, i.e. which address the language of R.

Mamma

Mamma returned 18 unrelated results before finally returning the following 19th value:

19. R: Testing Association of a Group of Genes with a Clinical Variable
checkerboard Checkerboard plot for Global Test ... exampleY Example (simulated) clinical variable for package "globaltest".
http://www.ugrad.stat.ubc.ca/R/library/globaltest/html/00Index.ht [Ask.com]

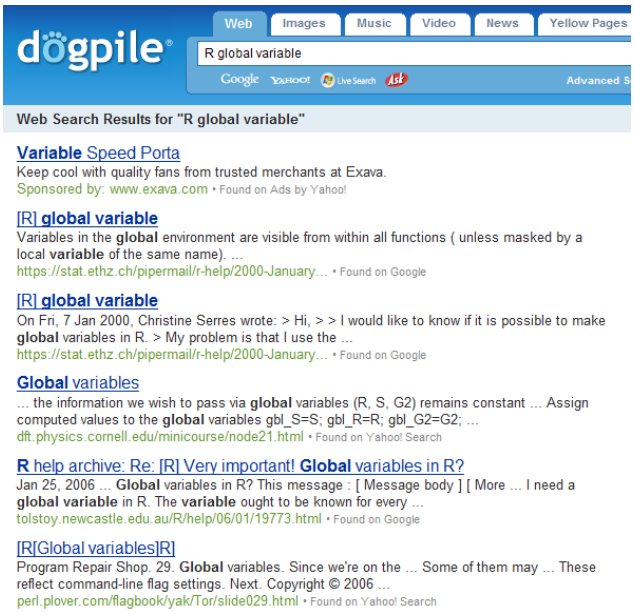
Quadsearch

QuadSearch impressively returned 90% of the Top 10 results as documents related to R. It actually wasn't until the 10th result that it returned an unrelated item.

Dogpile

Dogpile returned 5 pertinent results of their top 10 (50%).

Note: Dogpile's results looked as follows:



Because it is so difficult to distinguish the top result from any of the others, I consider it as the number one result. It does mention that it is "Sponsored by" someone, but because it does not set the result apart from any of the others at all in terms of its format, I will consider it as a normal result.

Objective Review

Timing

I used a Firefox extension to measure the page load times of each search engine. It seems that Mamma was inconsistently fast (but due to two of ten very slow load times, its mean came up above that of Dogpile). Quadsearch was consistently slower, and Dogpile was consistently average. When all three means are normalized over the lowest, you get the following results:

Quadsearch :	1.165
Mamma :	1.019
Dogpile:	1.000

Thus, you can observe that there wasn't any great disparity overall. This could be due entirely to bandwidth issues on their end or on the machine I was using. Or it could be a product of loading external files (I disabled images, but external CSS files or Javascript classes might still have been loaded). Also, it might have been a product of the time it took my browser to parse the page. One other consideration of this timing data is the potential that my results might have been cached, or stored in increasingly-accessible forms as I searched on it multiple times in a short period.

	Page Load Time (s)		
Trial	QuadSearch	Mamma	Dogpile
1	1.802	1.130	1.528
2	1.584	1.113	1.409
3	1.657	0.868	1.455
4	1.804	*3.670*	1.653
5	1.596	1.137	1.179
6	1.593	0.879	1.229
7	1.646	0.948	1.436
8	1.666	*3.520*	1.397
9	1.761	0.963	1.565
10	1.529	0.321	1.419
Mean	**1.6638**	**1.4549**	**1.427**
Median	**1.652**	**1.038**	**1.428**
StndDev	**0.10**	**1.15**	**0.14**

Precision and Recall

Information Retrieval is an inherently fuzzy area when it comes to success rates, false positives, etc. This is because what's precise for one user may not be precise for another, even when using the same query. Thus, even when using the most objective metrics possible, I'm still biasing these results on the basis of my intended outcome. If a user had been searching for a write-up on global variables, one of which was named R, it's quite likely that they may have ranked the three search engines in the inverted order as I will. This being said…

I will define the two metrics as follows:

Precision = |Relevant and Retrieved| / |Retrieved|

and Recall = |Relevant and Retrieved| / |Relevant|

I chose to analyze only the top 10 results, however I believe that 0/10 metric for Mamma was problematic and unrealistic, so I expanded its search to 19, so I could find one relevant result.

Because of the specific domain to which the query applies, it is a bit difficult to define the set of all relevant documents. However, I will assume that the domain-specific searching service I have used up to this point (rseek.org) is sufficiently large so as to serve as my gold standard of relevant documents. I queried "global variable" (as the site is only searching documents and sites specific to R, there was no need for the "R") and was given 71 results. I will use this value to represent the set of all relevant documents available on the Internet. I don't think it's unreasonable to assume that the true set is on the order of 71 documents.

	Dogpile	Mamma	QuadSearch
Relevant	5	1	9
Retrieved	10	19	10
Precision	**.5**	**.05**	**.9**
Recall	**.07**	**.01**	**.13**

Further research could be done to ensure that all of the discovered relevant documents by the three metasearch engines do, in fact, exist in the "master set" of all 71 documents found through rseek. I was unable to find a feasible way to export any of these lists, so I did not pursue this consideration further.

Summary

Overall, it seems that QuadSearch was, by far, the best fit for my query. This is interesting as it might represent that some of the latest research in metasearch engines really do have practical applications in searching.

Dogpile was not too far behind, however it did not perform as well as QuadSearch in any metric other than time.

Mamma was inferior in the most significant metrics - relevant and retrieved - though it did compete (80% of the time) in the time-trials of all the engines.

Works Cited

Akritidis, L., Katsaros, D., & Bozanis, P. (2008). Effective Ranking Fusion Methods for Personalized Metasearch Engines. *Panhellenic Conference on Informatics* , 39-43.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

http://quadsearch.csd.auth.gr/

www.rseek.org

www.dogpile.com

www.mamma.com