Library Philosophy and Practice 2008

ISSN 1522-0222

Overlap in Web Search Results: A Study of Five Search Engines

Rafiq Ahmad Rather
Dept. of Education
Govt. of Jammu and Kashmir, India

Fayaz Ahmad Lone
Documentation Officer
Centre of Central Asian Studies
University of Kashmir, India

Gulam Jeelani Shah
Professional Assistant
University of Kashmir, India

Introduction

The web is expanding exponentially. In January 2007, there were nearly 30 million pages (WWW FAQ, 2007). This expansion has led to reliance on search engines to find web resources. This in turn casts responsibility on the search engines to meet the needs and expectations of the scholarly community. Using more than one search engine is futile if overlapping is frequent and substantial. Overlapping is genuine if the common results are highly relevant to the user's query. Use of different search engines simultaneously reduces searching time and increases efficiency. Though search engines index multiple and separate resources, some results occur in many search engine's databases and in some cases a search engine retrieves results by indexing other search engines' databases. The present study is an attempt to identify search engines with less overlapping for use by the scholarly community.

Overlap Studies

In the ocean of literature on search engines features, precision, recall, and other technical aspects, there has been little attempt to study overlap. Bharat and Border (1998) measured overlap among websites indexed by Hotbot, Altavista, Excite, and Infoseek using 10,000 queries carried out at two different intervals of time in June 1999 and November 1999, and found that the overlap was very small, less than 1.4 percent of the total coverage. Ding and Marchionini (1998) evaluated results retrieved by Infoseek, Lycos, and Opentext to measure the level of common results and report a low level of overlap. Chignell, Gwizdka, and Bonder (1999) found little overlap in the results returned by various search engines and describe meta-search engines as useful. Gordan and Pathak (1999) studied five search engines by measuring overlap at a document cutoff value of 20, 50, 100, and 200 and find that approximately 93 percent of the results were retrieved by only one search engine. Nicholson (2000) replicated the 1998 Ding and Marchionini study and found similar results with low web search engine overlap. Ferrara, da Silva, and Delgado (2004) evaluated previous overlap studies with the finding that documents retrieved by multiple information retrieval systems in relation to the same query are more likely to be relevant. Spink, Jansen, Kathuria, and Koshman (2006) examined the overlap among results retrieved by three major web search engines (Google, Ask Jeeves, and Yahoo) using a set of 10,316

randomly selected queries. The study shows that the percentage of total results unique to only one of the three search engines was 85 percent, with 12 percent found by two of the three search engines, and 3 percent found across all three.

Scope of the Study

The study uses five search engines (Altavista, Google, Hotbot, Scirus, and Bioweb), of which first three are general and the last two pertaining to science and technology and biotechnology respectively. The study is further limited to the field of biotechnology for which search terms were extracted from LC List of Subject Headings (Library of congress, 2003).

Objective

The study measures the overlap among the search engine results to identify search engines with less overlap.

Method

The study was carried out in three stages: literature review, selection of search engines, and invention of queries.

Population Selection

One hundred fifty search terms were drawn from an international vocabulary tool (Library of Congress, 2003), then refined to twenty queries and grouped under simple, compound and complex queries.

Test Environment

Each term was submitted to the selected search engines in turn, using the basic or simple search. One query was searched each day using all five search engines. The first ten results were recorded and evaluated to determine common results. The results were also evaluated by their contents to avoid any possibility of occurrence of results under different URLs.

Measuring Overlap

The overlap between or among the select search engines is the set of results retrieved by each engine for a query and is represented by intersection (n). The names of search engines are abbreviated by the first letter. For the sake of convenience, “G n A exactly” is the set of results retrieved by Google and Altavista and not by any other search engine, and “G n A n H exactly” is the set of results retrieved by Google, Altvista, and Hotbot, and not by Scirus and Bioweb. The sets of results retrieved by each search engine separately are also reported.

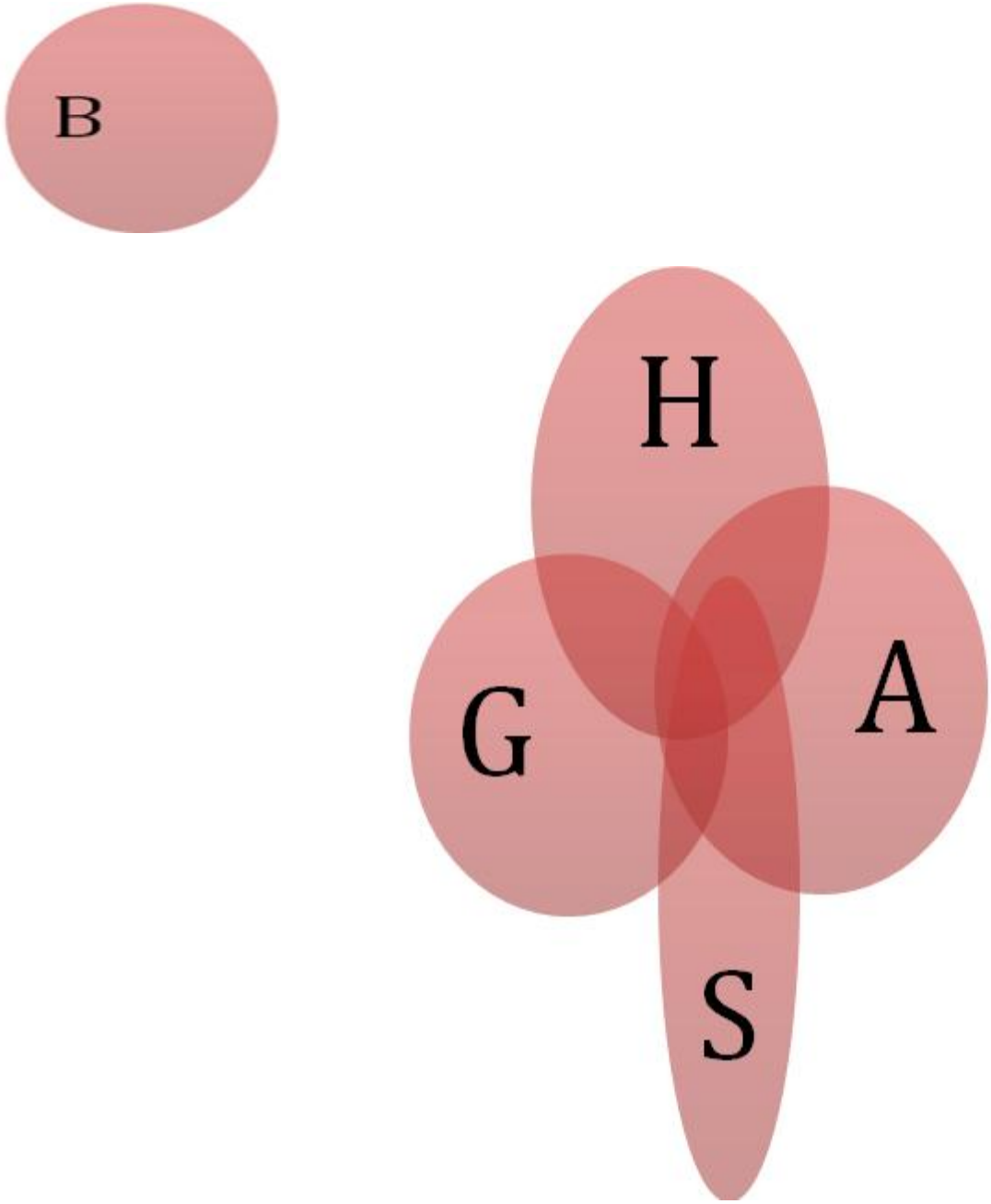
Results and Discussion

Analysis of results (Table 1) reveals that overlap is comparatively greater between Altavista and Hotbot (A n H), followed by Google and Hotbot (G n H), and Hotbot and Scirus (H n S). Overlap is considerable in Google, Altavista, Hotbot (G n A n H), followed by Google, Altavista, Scirus (G n A n S), while there is no overlap between Bioweb and other search engines (Figure 1).

Table 1

SET	No. of Results	SET	No. of Results
G exactly	166	G n A n H	007
A exactly	167	G n A n S	005
H exactly	170	G n A n B	000
S exactly	164	G n H n S	003
B exactly	200	G n H n B	000
G n A	007	G S nn B	000
G n H	010	A n H n S	004
G n S	007	A n H n B	000
G n B	000	A n S n B	000
A n H	011	H n S n B	000
A n S	006	G n A n H n S	002
A n B	000	G n A n H n B	000
H n S	008	A n H n S n B	000
H n B	000	G n A n H n S n B	000
S n B	000		

Figure. 1. Overlap



Bioweb retrieved 100 percent unique URLs, followed by Scirus (94.25 percent) , Altavista (92.26), and Google (91.21) (Figure 2). Hotbot has the highest degree of overlap (15 percent), followed by Google (8.79 percent) and Altavista (7.74 percent) (Table 2).

Figure 2. Percentage of Unique URLs

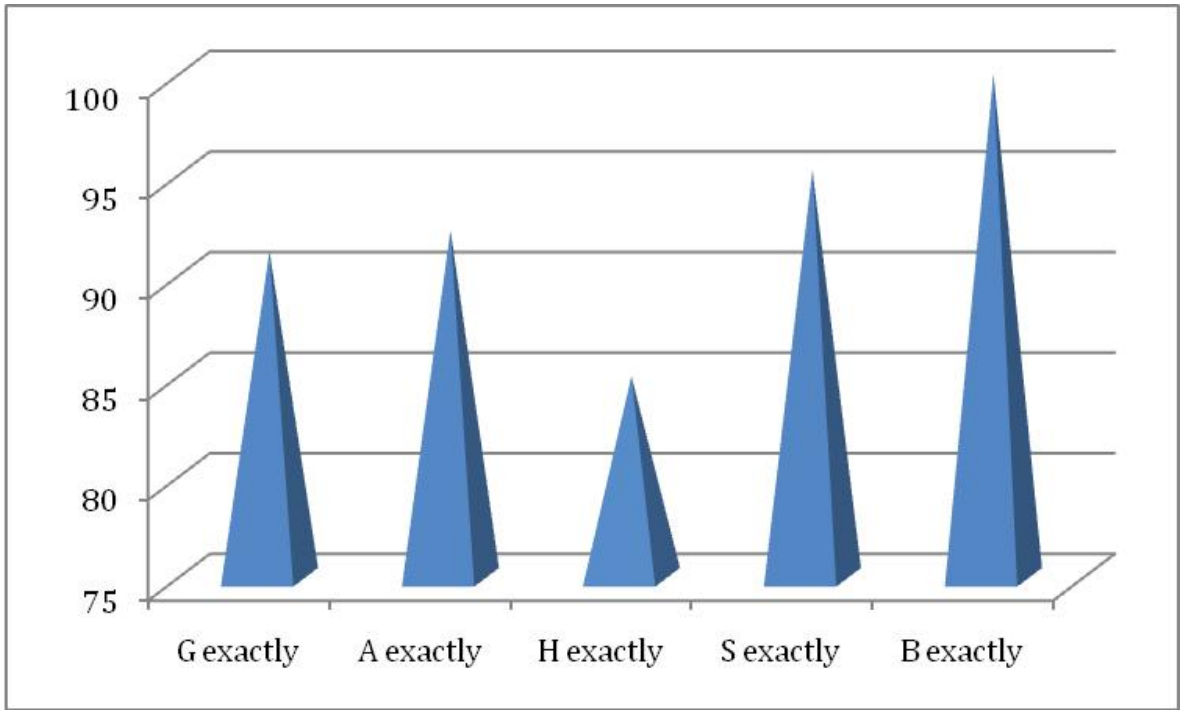


Table 2: Degree of overlap

Search Engine	Total URLs	Unique URLs	Degree of Overlap (percent)
Google	182	166	8.79
Altavista	181	167	7.74
Hotbot	200	170	15
Scirus	174	164	5.75
Bioweb	200	200	0.0

The nature of the queries influences overlap, which is more frequent in multiword (i.e., compound and complex) queries rather than one word queries (i.e., simple queries). There was no overlap in four of the simple queries, while all the compound and complex queries produced some overlap between or among the search engines. This analysis reveals that 92.53 percent of the URLs are retrieved by one search engine only (which could be any of the five), 5.22 percent are shared by two, while 2.02 percent and 0.21 percent of the URLs were retrieved by three and four search engines respectively.

The degree of overlap found is low in relation to previous studies (Nicholson, 2000 and Hord and Wilson, 2001) despite database growth. The overlap results are found to be relevant to an earlier study (Ferreira, da Silva and Delgado, 2004). Nevertheless, the overlap is not useful for simultaneous use of search engines in reducing searching time for users. Among the selected search engines, Hotbot had the most overlap (followed by Google) with other search engines except Bioweb. The reason for the overlap is the large database size of the search engine. This is evident, since Bioweb has no overlap with other search engines, and has a small and unique database. On the other hand, Bioweb does not come up to expectations because of its low precision and recall (Shafi and Rather, 2005) which do not keep up with the ever increasing growth of the web. The findings of the present study may not remain valid for long time due to the dynamic nature of the search engines.

References

Bharat, K., & Broder, A. (1998). A technique for measuring the relative size and overlap of public Web search engines. *Computer Networks and ISDN Systems 30* (1–7), 379–388.

Ding, W., and Marchionini, G. (1998). A comparative study of Web search service performance. In *Proceedings of the annual conference of The American Society for Information Science.* pp 136–142.

Chignell, M. H., Gwizdka, J., & Bodner, R. C. (1999). Discriminating meta-search: A framework for evaluation. *Information Processing and Management 35* : 337–362.

Gordon, M., & Pathak, P. (1999). Finding information on the World Wide Web: The retrieval effectiveness of search engines. *Information Processing and Management 35* : 141–180.

Nicholson, S. (2000). Raising reliability of Web search tool research through replication and chaos theory. *Journal of the American Society for Information Science 51* (8): 724–729.

Hood, W. W., & Wilson, C. S. (2001). Overlap in bibliographic databases. *Journal of the American Society for Information Science and Technology 54* (12): 1091–1103.

Library of Congress (2003). *Library of Congress subject headings* (volumes 1-5). Washington: Library of Congress Cataloging Distribution Service.

Ferreira, J., da Silva, A. R., & Delgado, J. (2004). Does overlap mean relevance? In *Proceedings of WWW/Internet 2004 (LADIS) conference.* Madrid: LADIS: International Association for Development of the Information Society.

Shafi, S. M., & Rather, R. A. (2005). Precision and recall of five search engines for retrieval of scholarly information in the field of biotechnology. *Webology 2* (2). Available: http://www.webology.ir/2005/v2n2/a12.html

Spink, A., Jansen, B. J., Kathuria,V., & Koshman, S. (2006). Overlap among major web search engines. *Internet Research: Electronic Networking Applications and Policy 16* (4): 419-426.

WWW FAQs (2007). How many websites are there? Available: http://www.bouteii.com/newfaq/misc/sizeofweb.html .