IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP5256

This paper appears in *Managing Modern Organizations Through Information Technology*, Proceedings of the 2005 Information Resources Management Association International Conference, edited by Mehdi Khosrow-Pour. Copyright 2005, Idea Group Inc.

Search Engine Result Bias: An Empirical Investigation of Commercial Web Based Search Tools

Kholekile Gwebu and Jing Wang

Dept of Mgt. & Info. Systems, College of Business Admin., Kent State University, Kent, OH 44242, USA, {kgwebu, jwang2@kent.edu}

INTRODUCTION

In recent years the world wide web has gained prominence as a prime resource for information on an array of topics including air travel, real estate and home décor just to name a few, which both individuals and organizations use to make informed decisions. Because the web is so vast and contains both structured and unstructured documents, web users often turn to web-based Information Retrieval Systems (IRS), typically referred to as search engines, as the main means of searching, sorting and navigating through the web. IRS in general have been in existence for decades and have allowed users to sort and search through structured documents, such as library records and news paper etc. IRS research tends to focus on the performance in terms of coverage, relevance, and ranking[1-6]. One major issue which has largely been ignored by researchers is that of bias in search engines. Bias simply refers to "undue inclusion or exclusion of certain items among those retrieved in response to queries or it is revealed in giving undue prominence to some items at the expense of others"[7]. Bias was previously never a serious issue in traditional IRS because the information being retrieved from them was not subject to systematic manipulation since it was largely non-commercial in nature. Today however, the competitive and commercial nature of search engines on the Web makes them vulnerable to systematic manipulation of results.

Only a handful of studies devoted to assessing search engine bias are available on leading scholarly research databases [7], and even in such articles the authors have called for additional research into this area. If search results from leading search engines are indeed systematically skewed, web searchers need to be extremely cautious when attempting to retrieve fair and unbiased information from the web as some relevant search results obtained from search engines could intentionally be substituted with irrelevant but more commercially or politically appropriate results by search engine companies. This paper investigates the nature and extent of bias in commercial search engines. We consider the most popular search engines for assessment, as they are the ones which tend to have the most impact on internet users, then we use over 200 real user generated queries to assess bias across 8 different subject areas for all the search engines.

The remainder of the paper is arranged as follows.The subsequent section synthesizes relevant literature on search engine bias. Thereafter, a set of hypotheses are presented followed by a description of the experiment conducted to assess bias, the empirical findings, and a discussion on those findings. Finally, we conclude by pointing out limitations of the study and issues which future researchers may wish to explore.

LITERATURE REVIEW

Search engine bias, an important search engine performance measure, has received little attention in traditional IRS literature. With web infrastructure becoming more robust, web information retrieval is becoming an increasingly important part of everyday life. As users' dependence on search engines grows, so too will their need for unbiased information [7, 8]. It is therefore imperative for IR researchers to assess and perform studies on search engine bias hence enhance users' knowledge on the nature of bias in various commercial search engine results and assist them in making informed decision on search engine choice. [8, 9].

Nevertheless, one major challenge remains in assessing bias in search engines: defining and measuring search engine bias. While content is typically the focus of analysis in detecting bias in a message, advertisement, and political propaganda [10, 11], identifying bias in search engines is different. As Mowshowitz et al.[7] point out, retrieval systems such as search engines produce a collection of items (including titles, citations, or brief description) in response to queries. Bias in search engines is exhibited in the selection of items rather than in the content of any message. Hence, bias in search engines is defined by Mowshowitz et al. [7, 8] as a function of emphasis and prominence. To be more specific, when a search engine gives undue prominence to certain items at the expense of others or places undue emphasis on certain items, the retrieved results are considered biased. In contrast, an unbiased system should produce a balanced and representative list of items from its database for any set of queries. Measurement of bias can therefore be operationalized as measuring the degree to which the distribution of items in a particular search engine's results deviates from that balanced and representative norm.

In their studies, Mowshowitz et al. [8] propose using a family of comparable search engines and computing the frequencies of occurrence of the URLs in the collection retrieved by the chosen search engines to approximate that ideal, balanced, and representative norm for a set of queries. A software system is also developed by the authors to facilitate empirical investigation of the applicability and utility of the measurement approach proposed in their studies. This software system acts as a meta-search engine and is able to invoke 15 commercially available search engines and automatically computes the bias value for any set of queries.

Studies conducted by Mowshowitz et al. [8] represent an advance in evaluation of search engine bias. But their studies exhibit some limitations. For instance, the subject domains and the search terms used to represent the subject areas were derived from the ACM Computing Classification System. This methodology of subject domain and key words selection exhibits one inherent problem. The chosen subject areas and key words could be too closely related to each other due to the fact that they all belong to the same super category: Computing. One could raise the issue as to "to what extent we could generalize the finding that search engine bias does not depend on subject domains searched. In other words, if very distinct rather than related subject domains and key terms are used, will the same results be obtained and the same conclusions be reached? Furthermore, the limited number of subject domains used in their studies and the limited test results also restrict us from generalizing the results found in their studies. The authors have recognized this limitation and called for more extensive testing and statistic analysis before any conclusions can be drawn about the relative performance of the studied search engines.

Finally, while ANOVA was used in their study, the assumptions of ANOVA have never been addressed in either of their studies. In addition, while the authors found differences in bias values between search engines are statistically significant in each of the subject area, no statistical analysis has been conducted on how search engines should be ranked based on their relative performance in terms of bias.

HYPOTHESES DEVELOPMENT

Our hypotheses are derived from the unexplored gaps in prior studies. The first hypothesis stems from the belief that different search engines are skewed from the norm in the results they display following a query, with some being more skewed/biased than other.

H1. Different search engines exhibit different levels of bias across a broad range of subject areas even if result ranking is not taken into consideration.

When result ranking is factored in we expect to see greater levels of bias in the results because search engines will tend to place more emphasis on results in which they have some vested interest than in those which they do not.

H2. Different search engines exhibit different levels of bias across a broad range of subject areas if result ranking is factored in.

H3a. There is a difference between ranked and non-ranked bias values across all search engines.

We expect that that when URL ranking is factored in, there should be a greater deviation from the norm in the bias values for most search engines, thus higher bias values on average.

H3b. The bias values which factor in ranking are likely to be greater than those that do not factor in ranking.

Next we consider bias within each search engine. Our goal is to compare bias values across different subject areas. We believe that individual search engines exhibit more bias towards particular subject areas than they do to others. This may be because they have some vested interest in a particular subject area. For instance if Search Engine A had many clients in the automotive industry who would like to advertise and very few clients in the home décor, we would expect the results of automotive related queries to be more biased than those in those of home décor.

H4. All search engines will exhibit more bias in certain subject areas than they do in others.

Our next hypothesis is based on groupings of various subject queries into subject categories. This we believe will help further highlight the difference between different subject areas. We expect that subject areas which are more commercial in nature are likely to have higher bias values than those less commercial in nature. This is because more commercial areas have higher levels of competition and in order to compete companies and organizations are likely to influence the results of query results.

H5a. There are statistically significant differences between various subject categories.

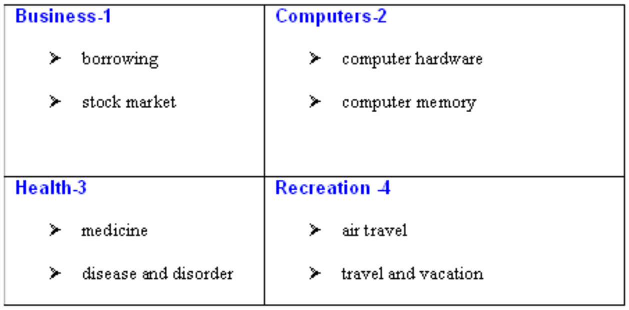
H5b. In subject categories where there is a high level of commercially oriented competition; there will be higher levels of bias than in subject categories which have less competition.

METHODOLOGY

Search Engine Selection

We selected the most prominent search engines for this study since they are more likely to reach a lager audience and have more clout on user decision making than the less popular ones. Additionally, larger search engines are more comparable in terms of results since they index comparable numbers of web pages. The ranking of search engine popularity used in this study was derived from Nielsen/NetRatings, a renowned internet and digital media audience analysis service. They provide web site ratings based on a sample of over 225,000 individuals (home and work surfers) who have real-time meters on their computers which monitor the sites they visit. The list of search engines we used was extended to cover two other very popular search engines namely, overture and Lycos. Therefore the entire list of search engines examined in the study consisted of Google, Yahoo, MSN, AOL, Teoma, Overture, and Lycos.

Table 1. Selected Categories and Subject Areas



Business-1	Computers-2
➢ borrowing ➢ stock market	➢ computer hardware ➢ computer memory
Health-3	**Recreation -4**
➢ medicine ➢ disease and disorder	➢ air travel ➢ travel and vacation

Subject Selection

The Open Directory Project (ODP) a comprehensive human-edited directory of various subjects on the web was used to select subject categories. Over time the 15 subject categories have emerged on (ODP). In order to ensure distinctness in the subjects areas selected for the study, each category was assigned a number from 1 – 15 then a random number generator was be used to select 4 categories. Thereafter two subject areas were randomly selected under each of the subject categories. The table below shows the eight different subject areas.

The search queries for each of the subject areas were obtained from Keywordcity.com which houses a categorized set of keywords and key phrases created by real users on the Goto.com search engine. The major categories and subcategories in which queries are grouped are inspired by those typically found on the Open Directory Project and the directory/ category pages of major search engines such as Yahoo, Lycos, AOL, and Google just to name a few.

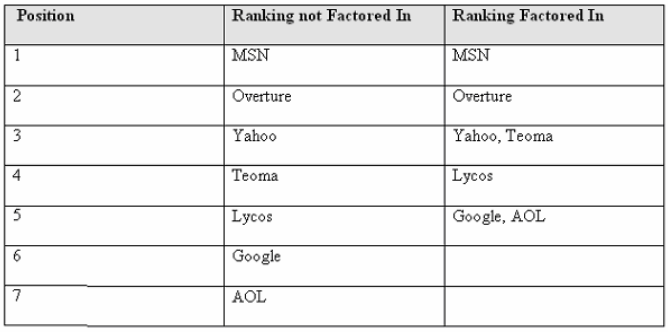
Keywordcity.com ranks keywords in terms of popularity. For each of the subject areas selected we chose the 15 of the most popular search terms. All the search terms in the Keywordcity database are not phrase searches i.e. they are not enclosed in quotation marks. To increase the number of search terms while retaining the underlying semantics of each original query we enclosed the 15 original queries in quotation marks resulting in a total of 30 queries for each subject area.

Instrumentation

The tool developed by [8] to assess the bias in search engines was used in the study. According to their documentation [13] the tool works as follows:

First a researcher specifies a collection of comparable search engines she wishes to compare. Search engines selected for the current study AOL, Google, MSN, Lycos, Teoma, Yahoo, and Overture were selected for investigation. This set of search engines is then used to estimate a fair or ideal distribution of items for a set of queries. Thereafter each of the search engine query results is compared to the ideal distribution and bias value (deviation from the ideal) is computed. The researcher may also factor in URL ranking, which involves weighting URLs which appear at the top of a list of retrieved URLs more than those appearing at the bottom. As with the computation of the previous bias value, each of the search engine query results is compared to the ideal distribution. In this

Table 2. Search Engine Bias Ranking



Position	Ranking not Factored In	Ranking Factored In
1	MSN	MSN
2	Overture	Overture
3	Yahoo	Yahoo, Teoma
4	Teoma	Lycos
5	Lycos	Google, AOL
6	Google	
7	AOL	

**Based on statistically significant differences (i.e. p<0.05)in multiple comparisons.*

study we ran two sets of queries; one that did not factor in ranking into the bias results and one that did. We were only interested in the bias values of the first 10 results as these are the ones which tend to have the greatest influence on people searching for information on the web.

STATISTICAL ANALYSIS AND RESULTS

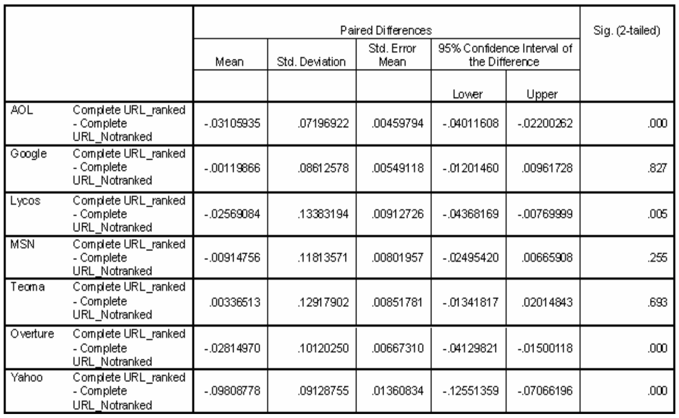
The first two hypotheses were assessed using ANOVA and were both supported (p-value of 0.000 in both cases) indicating that for each of the search engines selected, there were statistically significant differences in terms of bias when result ranking was not factored in and when result ranking was factored in. The assumptions of ANOVA i.e. each group is an independent random sample from a normal population, symmetric data and equal variances of population were first assessed. The first two assumption were satisfied but the homogeneity of variance assumption was not since the null hypothesis (the group variances are equal) was rejected in the Levene test. However ANOVA is robust to this violation when the groups are of equal or near equal size which is the case with the all the groups of queries we used in this study.

Using the Bonferroni multiple comparison procedure we were able to rank-order the search engines from most biased on the seven subject areas to least biased. The Bonferroni procedure "uses t-tests to perform pairwise comparisons between group means, but controls overall error rate by setting the error rate for each test to the experiment wise error rate divided by the total number of tests" (SPSS Manual 2004).The beauty of the Bonferroni procedure is that it adjusts the observed level of significance to cater for the simultaneous multiple comparisons thus minimizing type 1 error. Table 2 shows the result of this ranking. When ranking is factored in there is no statistically significant difference between Google and AOL and they both exhibit the lowest bias values. Additionally, Yahoo and Teoma are also not statically different from one another. In both cases MSN exhibits the highest bias values while AOL exhibits the lowest bias values.

For hypothesis 3a we compared the bias values with ranking factored in against those without. For this we aggregate all the bias values across subject areas and ran a paired samples t-test. A paired sample t-test is appropriate when one wishes to make a comparison of means of two scores obtained from two samples that are not independent as was the case in this portion of the study. The assumptions of this test i.e. 1.observations for each group should be made under the same conditions and 2. normally distributed mean differences (SPSS Manual 2004) were met.

The results of the paired sample t-test indicate a p-vale of 0.058 which is grater than 0.05 therefore we can not reject the null hypothesis that bias values which factor in ranking are significantly different to those that do not at the 95% level. This result is very close to being significant and therefore required further investigation. In order to understand why H3a was rejected we decided to separate search engines and look at whether ranked and non-ranked results were different.

Table 3. Paired Samples Test: Ranked vs. Non Ranked Bias Values for Each Search Engine



		Paired Differences					Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		
					Lower	Upper	
AOL	Complete URL_ranked - Complete URL_Notranked	-.03105935	.07196922	.00459794	-.04011608	-.02200262	.000
Google	Complete URL_ranked - Complete URL_Notranked	-.00119866	.08612578	.00549118	-.01201460	.00961728	.827
Lycos	Complete URL_ranked - Complete URL_Notranked	-.02569084	.13383194	.00912726	-.04368169	-.00769999	.005
MSN	Complete URL_ranked - Complete URL_Notranked	-.00914756	.11813571	.00801957	-.02495420	.00665908	.255
Teoma	Complete URL_ranked - Complete URL_Notranked	.00336513	.12917902	.00851781	-.01341817	.02014843	.693
Overture	Complete URL_ranked - Complete URL_Notranked	-.02814970	.10120250	.00667310	-.04129821	-.01500118	.000
Yahoo	Complete URL_ranked - Complete URL_Notranked	-.09808778	.09128755	.01360834	-.12551359	-.07066196	.000

It is evident from the above paired sample t-tests (Table 3) that AOL, Lycos, Overture and Yahoo, demonstrated a statistically significant difference between the ranked and non ranked bias values while the other three search engines did not. This may explain the statistically insignificant result we obtained when we aggregated search engines. Surprisingly, ranked results exhibited lower bias values than did non-ranked results in all statistically significant cases, thus counter to what we had hypothesized in H3b.

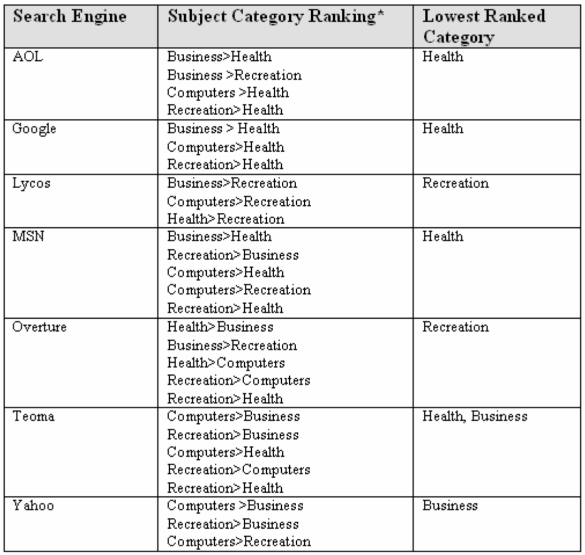
H4 involved analyzing bias values across different subject areas within each individual search engine. Take Google as an example, the null and alternative hypotheses are as follows:

Ho: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7 = \mu_8$

H1: At least one of the subject areas is not the same as the others

1 = borrowing, 2 = travel and vacation, 3 = stock market, 4 = computer memory, 5 = medicine, 6 = air travel, 7 = disease and disorder, 8 = computer hardware

Table 4. Subject Category Ranking



Search Engine	Subject Category Ranking*	Lowest Ranked Category
AOL	Business>Health Business >Recreation Computers >Health Recreation>Health	Health
Google	Business > Health Computers>Health Recreation>Health	Health
Lycos	Business>Recreation Computers>Recreation Health>Recreation	Recreation
MSN	Business>Health Recreation>Business Computers>Health Computers>Recreation Recreation>Health	Health
Overture	Health>Business Business>Recreation Health>Computers Recreation>Computers Recreation>Health	Recreation
Teoma	Computers>Business Recreation>Business Computers>Health Recreation>Computers Recreation>Health	Health, Business
Yahoo	Computers >Business Recreation>Business Computers>Recreation	Business

*Based on statistically significant differences in multiple comparisons.

Because H3a was rejected i.e. there is no statistical difference between the ranked and the non-ranked results we only consider the non-ranked results. ANOVA was used to test H4 for each of the seven search engines and it was found that there are significant differences in bias levels within each search engines across the 8 subject areas, thus H4 is supported.

Finally, when attempting to rank the subject areas from most biased to least biased in order to asses H5, we found it difficult to draw conclusive evidence as to which category is least biased. This could be because we aggregated search engines and while some search engines could be more commercially oriented than others, the difference would not be clear if all search engines were aggregated. Thus we separated search engines and ran multiple comparison test of the various subject areas test for each individual search engine.

From the above table it is clear that each search engine's subject categories differ in the level of bias , with the majority of search engines displaying the lowest bias levels in the health category, thus our last hypothesis although not supported does deserve further investigation. Future work would have to select categories which are clearly non-commercial and compare them to highly commercial ones in order to draw conclusive results.

IMPLICATIONS AND CONTRIBUTIONS

The paper addresses issues not fully addressed in previous Web-based IR research such as:

1. Whether or not there are statistically significant levels of bias within subject areas from one of the seven most popular search engines to another.
2. Whether or not the bias level across the results of the seven most popular search engines vary significantly from one subject area to another.
3. Whether or not bias between subject areas is significantly greater than the bias within subject areas.

Another contribution of the work done here has been the methodology employed. The assumptions of various statistical techniques used have been checked and adhered to, while in the limited previous similar studies, there is no mention of compliance to assumptions. Additionally, this study uses real life queries generated by real users, thus the study is not prone to bias which could be introduced by researchers who create their own set of queries as in previous studies.

These results serve to inform decision makers at the organizational and individual level of the dangers of only utilizing a single search engine when searching for information which could subsequently be used to make important decisions.

LIMITATIONS AND FUTURE RESEARCH

Like all studies, this work has some limitations. First, only 8 subject areas were selected for this study, future studies may select different areas to analyze. It may also be interesting to explore whether or not there is bias between search engines within subcategories of certain subject areas. With adequate resources, future researchers could consider substantially increasing the number of subject areas examined so as to cover a broad spectrum of areas. Results derived from such studies could be documented, published and tracked over time, giving users an idea of the extent of bias in a search engine performance over time.

Additionally, more search engines could be considered. In this study, only the seven most popular search engines were considered. However, there are other search engines in existence today. It would be interesting to know the extent of bias on these search engines.

Finally, future research could examine meta-search engines and whether or not there is bias in the results sorted according to their own relevance criteria. Thereafter comparisons could be done between the meta-search engines and individual search engines.

REFERENCES

Please contact the authors for a complete reference list.