BIAS ON THE WEB

When it comes to measuring bias on the Web, there is clearly strength in numbers (of search engines, that is).

YOUR REFRIGERATOR STARTS MAKING UNUSUAL NOISES; you figure it is time to replace it. Since the old one worked faithfully and reliably for some years, you have paid little attention to the marketplace for refrigerators, and apart from a few brand names are unfamiliar with replacement options. Unwilling to take the time to head for the local library and browse consumer magazines for "refrigerators," you decide to search the Web for information about this home appliance.

Obvious search terms include "refrigerator" and "refrigerators." The singular form may point to secondary uses of refrigerators, for example, as places to display children's art. It might also be used to designate what is kept in the appliance: Web pages dealing with the "healthy refrigerator" might be retrieved. The plural form is more likely to pick up Web pages dealing with products and services. To narrow the search and exclude commercial refrigerators, one might use the term "home refrigerators."





To illustrate differences between search engines, we executed a search last September using the term "home refrigerators." An experimental metasearch engine ("Piggy") developed by the authors and accessible at www-cs.engr.ccny.cuny.edu/~project/ was used to perform the search. (See [5] for definition of a meta-search engine.) Nine different, commercial search engines—Yahoo, Google, HotBot, Goto, AltaVista, Excite, Lycos, Northern Light, and LookSmart—were invoked in turn by the metasearch engine. The first 50 URLs retrieved by each of these engines were collected. Only 309 of the 450 URLs retrieved were distinct, representing 231 unique Web sites.

The Web pages retrieved were mainly those from Web sites of manufacturers, vendors, and organizations offering information about refrigerators. Significant differences in coverage of the manufacturers of refrigerators are apparent in the search results. If one relies exclusively on say, Excite, one may not notice offerings by GE and Maytag; if Google is the engine of choice, one might overlook Kitchenaid and Whirlpool. In both cases URLs corresponding to the manufacturers' Web sites were not found among those retrieved.

The differences can be illustrated by examining the appearance of brand names in the URL strings. Consider the following well-known brand names in the U.S. market: Amana, Frigidaire, GE, Kitchenaid, Maytag, and Whirlpool. Only one search engine retrieved a URL containing the name Amana. Different sets of four of the nine search engines retrieved URLs containing the names Kitchenaid and Whirlpool, respectively. Frigidaire, GE, and Maytag appear in different sets of five of the nine engines. Table 1 provides a complete list showing which search engines turned up which brand names among the URLs they retrieved.

Although the presence or absence of substrings designating brand names does not tell the whole story about differences between search engine results, other aspects of the Web pages reinforce the foregoing analysis. The observed differences reflect what may be termed "bias" in the sets of URLs retrieved by the engines.

Biased search results on product information illustrates a general problem of considerable social importance. The Web is replacing the traditional repositories that individuals and organizations turn to for the information needed to solve problems and make decisions. Search engines are gateways that mediate between users and the billions of pages on the Web, in essence acting as automated reference librarians. Users need to be aware of potential bias in search engine results to enable them either to seek alternative sources (for example, other search engines) or to hedge conclusions based on the results. This article presents an operational definition of search engine bias, describes and illustrates a system for measuring bias, and presents a statistical analysis designed to demonstrate the utility of the measure.

What is Bias?

Political scientists and students of advertising have long conducted research on bias. This research has given rise

to content analysis—a systematic method for detecting propaganda or bias in communications like speeches and media reports [1]. Scholars concerned with social aspects of computers have investigated bias in information systems. These studies focus on hidden assumptions or values that influence design, development, or use of such systems [2, 6, 7, 10].

Neither of these approaches is directly applicable to the measurement of bias in search results. A retrieval system such as a search engine contains a collection of items (typically titles, citations, or brief subject descriptions) that represent messages, rather than the messages themselves. Bias is exhibited in the selection of items, rather than in the content of any particular message. We call the former "indexical bias;" the latter, "content bias" [8].

Indexical bias in a set or list of URLs retrieved in response to a query is a function of emphasis and prominence. It is related to other "quality of information" issues, but captures an aspect of quality that differs from relevance, accuracy, timeliness, and so on. A collection of items retrieved from a database may exhibit bias *whether or not the items are relevant to a user's query.* The purport of the "whether or not" caveat is clear from some extreme cases. If the items retrieved are all deemed relevant by a user, there may be others—not retrieved—that would also be considered relevant by that user. On the other hand, a set of items irrelevant to one user might be relevant to another user for the very same query.

Given a norm prescribing expected frequency or prominence of items retrieved in response to a query, a set exhibits bias when some items occur more frequently or prominently with respect to the norm, while others occur less frequently or prominently with respect to the norm [2, 8]. The absence of certain brand names in the refrigerator example signifies bias in the results of a particular engine because other engines do retrieve those brand names. Prominence is reflected in the position a URL occupies in the list of items retrieved for a given search term. The norm used in the research reported here is based on the idea of pooling the results of a basket of search engines. This norm lends itself to a practical measurement scheme.

A Measure of Bias

As suggested, the bias exhibited by a set or list of URLs deals with emphasis, that is, the balance and representativeness of items in a collection retrieved from a database for a set of queries. This calls for assessing the degree to which the distribution of items in the collection deviates from the ideal or norm. A family of comparable search engines acting on a set of related search terms can be used to approximate the ideal or fair distribution of items retrieved for a set of queries. The distribution is obtained by computing the frequencies of occurrence of the URLs in the collection retrieved by several search engines for the given queries. Bias can be assessed by measuring the deviation from the ideal of the distribution produced by a particular search engine.

		Yahoo	Google	HotBot	Goto	Alta Vista	Excite	Lycos	Northern Light	Look Smart
Amana						x				
Conserv		x		x	x					
Explorer		x	x	x	x		x	x	x	
Frigidaire				x	x	x	x			x
GE		x	x	x	x			x		
Inglis		x	x	x	x	x				
Kitchenaid				x	x	x	x			
Klondike		x	x	x	x	x			x	
MAC-GRAY		x	x							
Maytag		x	x	x	x					x
Roper										x
Sun Frost		x	x				x			
Whirlpool				x	x	x	x			

Table 1. Brand names vs. search engines.

The pooling of results produced by searches on related terms approximates the collection of URLs that might be expected from the use of an ideal query. This provides a better basis for comparing the results of one engine with those generated by a group of engines. For example, the results of a search on "home refrigerators" could be augmented by those retrieved for "fridges," "ice boxes," "refrigerator," "refrigerators," as well as other terms that might be used to search for information on refrigerators for the home. Although duplicates occur occasionally, normally the

{ BY ABBE MOWSHOWITZ AND AKIRA KAWAGUCHI }

URLs retrieved by a search engine for a single query are all distinct, so the frequency of occurrence of any one of them would be zero or one. Using the pooled results from several independent searches on the same subject gives a range of frequency values between 0 and 1 and thus provides a more refined and accurate picture of the emphasis accorded a URL by a search engine.

Bias linked to excessive or insufficient prominence of items in the list of URLs retrieved by a search engine can be captured by weighting the frequency count of the URLs. That is to say, instead of incrementing by one, the contribution to the count can be modified to reflect the position occupied by the URL in the list of items retrieved.

Lacking a search engine that finds all and only relevant items, users would be well advised to take account of the bias characteristics of any particular search engine. If the results of tests for bias were regularly published on the Web, users would be in a better position to make informed choices about search engines. They might, for example, realize the need to use several engines rather than relying on just one. Metasearch engines offer a partial solution, although their bias performance depends on the particular engines they invoke.

The quantitative measure we are using in our research is not meant to capture every connotation associated with the word "bias" in ordinary discourse. However, preliminary applications show that the measure operationalizes the critical features of bias in a retrieval system, and that it can be used as a diagnostic instrument to help designers and users assess the objectivity of search engines as conveyors of information. The bias measure is based on a procedure for obtaining a collection of URLs corresponding to a set of queries processed by a set of search engines. The measurement and computational procedures described in detail in [8] are illustrated here.

The significance of indexical bias is especially clear for controversial topics. Thus, we have chosen to illustrate the computation of bias by searching the Web for information on "euthanasia." Three search engines (HotBot, Google, and Northern Light) and two search terms ("euthanasia" and "mercy killing") were used to collect URLs on the subject of euthanasia. Once again, we used the Piggy experimental metasearch engine.

To keep the illustration simple only the first 10 URLs retrieved by each search engine were collected. Thus the total number of URLs collected was 60 = 10 x 3 x 2: 10 produced by each of three engines for each of two search terms. The number of distinct URLs was 48, representing 44 distinct Web sites. This collection of 44 URLs, together with their frequencies of occurrence, is used in this example to define the norm for distribution of responses in the search for information on euthanasia. The URLs used in this example are truncated versions of the ones actually retrieved, that is, everything after the top-level domain name is eliminated. Thus, different Web pages attached to the same Web site are not distinguished.

The respective frequencies in descending order of the 44 different Web sites are shown in the following vector:

(7,3,2,2,2,2,2,2,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)

To compute the bias of a search engine, a corresponding vector of frequencies has to be determined. Google, for example, retrieves 10 URLs for each of the two search terms for a total of 20 URLs. Using the order of URLs in the pooled list for all three engines to order the components, we obtain the following frequency vector for Google:

(3,1,1,0,2,1,1,1,1,1,0,0,0,0,0,0,1,0,0,1,1,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,1,1,0,0)

To compute the bias of Google for this search, we first calculate the similarity of the two vectors using a well-known metric, to wit, their dot product divided by the square root of the product of their lengths [9]. Note that this value is the cosine of the angle between the two vectors passing through the origin of a 44-dimensional Euclidean space. The bias value is obtained by subtracting this from one, which again gives a number between 0 and 1. In our example, the similarity of the two vectors is given by:

$48/(124 \times 28)^{1/2} = 0.8146$
So, the bias value is 1 - 0.8146 = 0.1854

To take the position of a URL into account, we can modify the frequency count by using an increment value different from 1. One possibility is to increment a URL by $(m+1-i)/m$ where m (10 in this example) is the total number of positions and i is the position of the URL in the list. The bias value obtained using this increment is 0.1138. This value is smaller than the one computed when the increment is 1, but in general the bias value taking account of position could be larger or smaller.

The numbers of engines and search terms used in this example are too small to yield reliable bias values. However, the example does reveal the kind of information that can be gleaned from bias measurement. Using truncated URLs and including the engine's results in the computation gives the following:

HotBot 0.19037
Google 0.18539
Northern Light 0.52135

The relatively high value for Northern Light stems mainly from that engine's inclusion of proprietary Web pages among those retrieved. There is only a slight difference between HotBot and Google, suggesting the possibility that they share the same basic engine. However, if the URLs retrieved by the engine whose bias is being measured are excluded from the pool defining the norm, the difference in bias values increases. In the case of exclusion, we have:

HotBot 0.54293
Google 0.49492
Northern Light 0.93472

In general, a high bias value means the collection of URLs retrieved by search engine *A* deviates substantially from the norm. This should alert a potential user to anomalies in *A*'s performance. At the margin, search engine *A* may miss URLs that an ideal engine finds or *A* may find URLs the ideal engine does not notice. Bear in mind the URLs missed by *A* may or may not be relevant to the query, and the ones found by *A* that are passed over by the ideal engine also may or may not be relevant. However, if the ideal engine (that is, norm for the bias measurement) is defined in terms of a set of search engines that score highly on recall and precision (see [9]), it is probable the URLs missed by *A* are indeed relevant and the extraneous ones found by *A* are not relevant.

Experimental Results

To make effective use of the bias measure in assessing search engine performance we need to determine just what qualifications may apply to its use. Does the bias of an engine vary from one subject domain to another, for example, does an engine with a high bias rating on searches related to, say real estate, turn out to be relatively unbiased on searches concerning travel? Do the computed values depend on the search terms employed to represent a given subject domain? Are differences in bias between engines significant?

To resolve these questions and demonstrate the utility of the bias measure, we conducted a series of experiments last September for which the norm was defined by a collection of 15 commercial search engines. Subject domains and the keywords or search terms used to represent them were chosen from the ACM Computing Classification System. A classification scheme developed by subject experts simplifies the task of selecting representative search terms. (Details of the tree-structured ACM System are available at www.acm.org/class.) We chose eight subject areas under the three major headings B.7 Integrated Circuits, C.1 Processor Architectures, and D.2. Software Engineering. For each of the eight subject areas, the five (lowest level) subheadings were used as search terms. In each search session (that is, the processing of a set of search terms), the top 30 URLs returned by the 15 search engines for each of the terms were used to compute the bias values.

Table 2 shows the results of an analysis of variance of bias values across subject areas. The p-values in the table measure the credibility of null hypothesis which is usually rejected if the computed p-value is less than or equal to the widely accepted figure of 5% (0.05) error.

Except for AltaVista and MSN, the p-values are larger than 0.05. This means that, except for AltaVista and MSN, the null hypothesis cannot be rejected so that with these exceptions, the computed bias values are not sensitive to the choice of subjects areas.

The results of two analyses of variance are shown in Table 3. The p-values in the search engine row, all of which are 0.000, show the null hypothesis, must be rejected for each subject area. This means the differences in bias values between search engines are statistically significant in each of the subject areas examined. Figure 1 illustrates the separation between search engines on bias values computed for the keywords in the subject area D.2.4 (Software/Program Verification). Some of the variation in bias performance evident in the figure might be attributable to differences in the sizes of the search engines' indexes—bias might vary inversely with index size. Accounting for differences in bias performance is an open research question.

The p-values in the keyword selection row of Table 3 are all greater than 0.05, indicating the null hypothesis is to be accepted for all eight subjects. This means for any of the eight subject areas selected for this experiment, the distributions of the bias values computed for the 15 search engines exhibit no statistically significant differences, even though the search terms used to represent a given subject are different.

Conclusion

The statistical analyses described here show the bias measure discriminates between search engines, but for

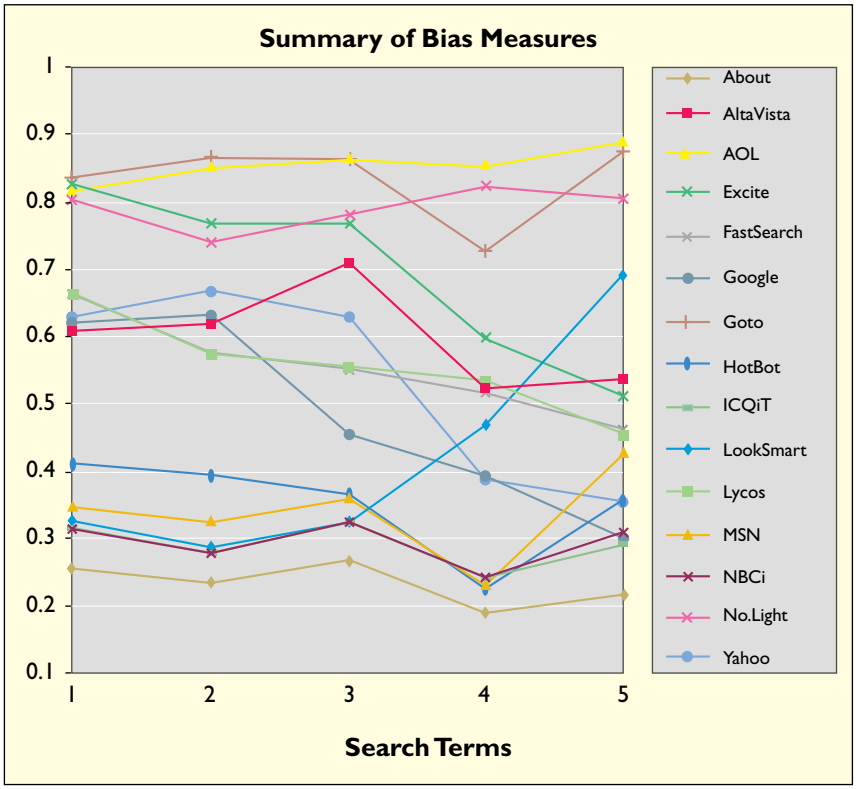

Figure 1. Bias values in subject D.2.4 for 15 engines.

most of the search engines bias does not depend on the subject domain searched (contrary to earlier expectations [8]), nor does it depend on the search terms chosen to represent the subject domain. These results support our contention that the measure of bias discussed here is a useful tool for assessing search engine performance. Similar results on a range of subject domains and search terms would justify using the bias measure to benchmark search engine performance.

Regardless of the utility of this particular measure, it is clear that bias on the Web is a socially significant issue [3, 8]. The only realistic way to counter the ill effects of search engine bias on the ever-expanding Web [4] is to make sure a number of alternative engines are available. Elimination of competition in the search engine business is just as problematic for a democratic society as consolidation in the news media. Both search engine companies and news media firms act as intermediaries between information sources and information seekers. Too few intermediaries spell trouble. C

Engine	About	AltaVista	AOL	Excite	Fast	Google	Goto	HotBot	ICQiT
p-value	0.070	0.028	0.260	0.144	0.111	0.057	0.544	0.174	0.486

LookSmart	Lycos	MSN	NBCi	NorthernLight	Yahoo
0.377	0.866	0.025	0.287	0.668	0.106

Table 2. Analysis of variance across subjects.

Subject	B.7.1 Types and Design Styles	B.7.2 Design Aids	B.7.3 Reliability and Testing	C.1.2 Multiple Data Stream Arch.	C.1.3 Other Arch. Styles	D.2.3 Coding Tools and Techniques	D.2.4 Software/ Program Verification	D.2.5 Testing and Debugging
Search Engine p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Keyword Selection p-value	0.101	0.507	0.505	0.709	0.890	0.230	0.716	0.407

Table 3. Analysis of variance across search engines and keyword sets.

References

1. Carley, K. Content analysis. *The Encyclopedia of Language and.Linguistics.* R.E. Asher, Ed. Pergamon Press, Edinburgh, 1990.
2. Friedman, B. and Nissenbaum, H. Bias in computer systems. *ACM Trans. Information Systems 14,* 3 (1996), 330–347.
3. Introna, L. and Nissenbaum, H. Defining the Web: The politics of search engines. *IEEE Computer 33*, 1 (2000), 54–62.
4. Lawrence, S. and Giles, C.L. Accessibility of information on the Web. *Nature 400* (July 8, 1999), 107–109.
5. Liu, J. Guide to meta-search engines. *BF Bulletin* (Special Libraries Association, Business and Finance Division) 107 (Winter 1998), 17–20.
6. Mowshowitz, A. *The Conquest of Will: Information Processing in Human Affairs.* Addison-Wesley, Reading, MA, 1976.
7. Mowshowitz, A. The bias of computer technology. In *Ethics and the Management of Computer Technology.* W.M. Hoffman and J.M. Moore, Eds. Oelgeschlager, Gunn & Hain, Boston, 1982.
8. Mowshowitz, A. and Kawaguchi, A. Assessing bias in search engines. *Information Processing and Management 38,* 1 (2002), 141–156.
9. Salton, G. and McGill, M.J. *Introduction to Modern Information Retrieval.* McGraw-Hill, New York, 1983.
10. Winner, L. Do artifacts have politics? *Daedalus 110,* 4 (Winter 1980), 121–136.

Abbe Mowshowitz (abbe@cs.ccny.cuny.edu) is a professor in the Department of Computer Science at The City College of New York. **Akira Kawaguchi** (akira@cs.ccny.cuny.edu) is an assistant professor in the Department of Computer Science at The City College of New York.