Goh, D.H., and Ang, R. (2002). Are pay for performance search engines relevant? *Journal of Information Science*, 28(5), 349-355.

Are Pay for Performance Search Engines Relevant?

Authors

Dion H. Goh
Division of Information Studies
School of Communication and Information
Nanyang Technological University
31 Nanyang Link
Singapore 637718
Singapore

Telephone: +65 6790 6290
Fax: +65 6794 0096
E-mail: ashlgoh@ntu.edu.sg

Rebecca P. Ang
Psychological Studies
National Institute of Education
Nanyang Technological University
1 Nanyang Walk
Singapore 637616
Singapore

Telephone: +65 6790 3215
Fax: +65 6896 9410
Email: phrang@nie.edu.sg

Are Pay for Performance Search Engines Relevant?

Abstract

Pay for performance (PFP) search engines, like their "traditional" counterparts (e.g. Google), provide search services for documents on the World Wide Web. These search engines however rank documents not on content characteristics but according to the amount of money a vendor is willing to pay when a user visits a Web site appearing in the search results page. A study was conducted to compare the retrieval effectiveness of Overture (formerly GoTo, a PFP search engine) and Google (a traditional search engine) from an academic perspective. Thirty-one queries from different graduate-level subject areas were submitted to each of these search services and the first 20 documents returned were retrieved and analyzed for precision and distribution of relevant documents using different relevancy criteria. Results indicate that Google outperformed Overture in both categories. Implications for this study are also discussed.

1. Introduction

Search engines are among the most heavily used online services on the World Wide Web. In Jupiter Media Metrix's recent list of top-20 web and digital properties for example, three sites that provide search and/or directory services ranked among the top-five [1]. This phenomenon can no doubt be attributed to the glut of information on the Web.

Popular search engines such as Google, AltaVista, Lycos and Excite assist users in filtering this glut for relevant web pages through the concept of relevancy ranking. Here, search results are sorted according to ranking algorithms for determining how closely a

document matches a query. The criteria used by such algorithms vary, and may include number of matching terms, frequency of terms, location of terms within the document and so on.

While search engines provide an invaluable service to users, they are ultimately owned by for-profit organizations. Business models may vary among these companies but most rely on advertising to generate a significant part of their revenue. Typical forms of advertising products offered include targeted banner advertisements, "featured" links that appear at prominent locations of the search results page [2], and pop-up windows [3] in which specific advertisements are displayed depending on the query terms entered.

The burst of the technology bubble in 2000 and the ensuing slump in the world economy however have caused search engine and other "new economy" companies to shift away from advertising-reliant business models because of cost-cutting measures initiated by their clients. Amid falling revenues experienced by these companies, one segment of this business has managed to thrive – the pay for performance search engines.

Pay for performance (PFP) search engines, like their "traditional" counterparts (e.g. Google and AltaVista), provide search services that rank retrieved web pages according to their similarity with a user's query. Unlike their traditional counterparts which employ ranking algorithms based on web page characteristics, PFP search engines rank documents according to the amount of money bid for certain keywords. The owner (or other interested parties) of the Web site registers with the search engine and determines appropriate keywords for it and the amount of money he/she is willing to pay when a user visits the site when it appears in the search results listings. Higher bids will mean a higher

relevancy ranking for the site, and hence a greater likelihood that a user will visit that Web site when a search is executed. Two of the more popular PFP search engines are Overture (formerly named GoTo) and FindWhat.com.

1.1 Rationale for Study

The results from PFP search engines may be comparable to advertising products (such as banner advertisements) which determine the type of advertisements to display given a user's search terms. The major difference is that PFP search engines present these advertisements as relevant web pages to a user's query. With click-through rates for banner advertisements at less at 1%, PFP search engines thus appear to be better positioned to meet the needs of vendors because of this "repackaging".

While PFP search engines may be ideal for vendors, do they meet the information needs of users? Do PFP search engines produce biased search results that exclude relevant information for a particular query? The present study seeks to address these issues from an academic perspective and attempts to determine if students can rely on PFP search engines to retrieve information relevant to their educational needs. Two search engines – Google (representing a traditional search engine) and Overture (representing a PFP search engine) were employed in this study. Google was used because of its popularity among searchers on the Web while Overture was chosen because of its dominant position among PFP search engines.

Thirty-one queries from different graduate-level subject areas taught at the Division of Information Studies at Nanyang Technological University were submitted to each of these services and the first 20 documents returned were retrieved and examined for

various factors such as number of relevant documents and distribution of relevant documents. These results then formed the basis for evaluating the effectiveness of PFP search engines for satisfying information needs in educational settings.

2. Background

Reactions towards the PFP search engines have been mixed, and comparisons of their performance against traditional search engines are mostly anecdotal. This section briefly discusses how Google and Overture, the search engines used in this study, operate, highlights the controversies involving PFP, and surveys existing work in search engine evaluation.

2.1 The Search Engines Under Study

Google currently offers the largest index of documents (more than 1 billion) [4] among search engines, and has quickly surpassed established services such as AltaVista in popularity. The search engine departs from traditional measures of relevancy that use variations of term-weighting algorithms, and instead measures relevancy by analyzing the link structure of the Web [5], an approach borrowed from citation analysis. The measure, known as PageRank, determines a page's citation importance or quality by calculating the number of Web pages that link to a particular page as well as the quality of those pages (their respective PageRanks). In essence, a higher PageRank value would thus indicate that a Web page is more relevant to a query because other important Web pages link to it.

Founded in 1998, Overture (formally named GoTo) is currently the market leader among PFP search engines. Overture ranks Web pages not through content analysis but by the

amount of money paid for keywords. In this approach, vendors anticipate the query terms that searchers would employ and bid for these terms, with the amount (minimum $0.05) indicating how much the vendor pays Overture whenever a searcher visits the Web page [6]. Higher bid amounts result in Web pages being ranked higher in the search results listings. To prevent compromising the performance of the search engine, Overture enforces various relevancy guidelines so that vendors may bid on a query term only if the Web page has "substantial content that is clearly and obviously reflective of the search term, and the line listing (title and description) accurately describes why the Web site is listed for the search term" [7]. In addition to displaying Web pages from its own index, Overture also supplements search results with unpaid listings from Inktomi. As expected, these unpaid listings appear only after Overture's paid listings, with paid listings distinguished from unpaid ones by a "Cost to advertiser" label next to each returned item.

2.2 The PFP Controversy

Festa [8] notes that the least popular vendors would typically bid higher amounts for keywords in PFP search engines in an attempt to be ranked highly for a given query. Consequently, these vendors would appear at or near the top of the search results page and users are likely to get documents that are opposite of what they were looking for. Likewise, Brin and Page [5] argue that advertising funded search engines are biased towards advertisers and against the needs of users. Using OpenText, an early PFP search engine that sold vendors the right to be listed at the top of search results for particular queries, as an example, the authors contend that "this type of bias is much more insidious than advertising" because it is not clear to users if certain pages are ranked according to content or according to the amount of money paid. This view is also shared by Marchiori

[9] who argues that such "preferred listings" result in performance degradation of search engines.

Sullivan [10] however, provides a more neutral view, comparing PFP results to newspaper advertisements and Yellow Pages listings. He argues that just as these various forms of advertisements provide useful information to the reader, PFP search engines may provide useful results that traditional search engines may not return, and hence users might not have otherwise seen. Similarly, Kopytoff [11] interviewed executives from several Internet-based companies and found that all agreed that paid listings are useful, especially with helping people buy products online. In addition, these executives claim that the problem of "spamming", in which irrelevant keywords are used to increase a document's relevancy ranking, is reduced because of search engine policies as well as cost to vendors.

On the other hand, Sullivan [10] cautions against PFP search engines providing only paid listings because users require a wide variety of information, not only those from vendors who can afford to advertise. In addition, Sullivan argues that search engines should also provide greater disclosure and inform users that their search results depend on the PFP concept. Doing so will help users make better decisions during the search process.

The failure of search engines to adequately inform users that paid listings are returned within search results has led a complaint being filed to the US Federal Trade Commission by Consumer Alert, a United States-based consumer group [12]. The complaint claims that eight companies that provide search services - AltaVista, AOL Time Warner, Direct Hit, iWon, LookSmart, Microsoft and Terra Lycos, are violating US law by "placing ads

in search engine results without clear disclosure that the ads are ads". In addition, the complaint states that these listings "look like information from an objective database selected by an objective algorithm. But really they are paid ads in disguise." It should be noted that many of the search engines named in the complaint obtain their paid listings from Overture. The company itself was not named in the complaint probably because it provided a "Cost to advertiser" label next to each returned item to indicate that it was an advertisement.

2.3 Search Engine Studies

Several studies evaluating search engine performance in academic settings have been conducted and this section highlights some of them. Chu and Rosenthal [13] used queries extracted from reference questions handled by librarians at Long Island University to evaluate precision and response times of three search engines (AltaVista, Excite and Lycos). The questions covered a variety of topics that included philosophy, neurobiology and psychology. Each query was submitted to the three search engines and precision values for the first 10 documents returned for each query as well as the mean precision for the 10 queries were calculated. Results indicated that AltaVista outperformed Excite and Lycos in terms of precision, although response time differences were not significant.

Similarly, Leighton and Srivastava [14] compared the performance of five search engines (AltaVista, Excite, Hotbot, Infoseek and Lycos) using actual questions asked by undergraduate students at a university library reference desk. Employing five different measures for relevance, precision values for the first 20 documents (*first 20 precision*) returned were calculated. The authors found varying performance among search engines,

although AltaVista, Excite and Infoseek generally produced better precision values than the other search engines.

Finally, Gordon and Pathak [15] adopted a different approach and employed expert searchers to seek information on the Web for 36 faculty members at a university business school. Topics were varied and included accounting, law, information systems and international business Searchers were instructed to perform queries for each information need repeatedly until the most relevant documents among the first 200 were retrieved. The search engines that provided the highest precision rates were AltaVista, Open Text and Lycos.

3. Methodology

Several experiments were conducted to compare the performance of a traditional search engine (Google) and a PFP search engine (Overture) in meeting academic information needs. As discussed earlier, these search engines were chosen because, excluding portal/directory sites such as Yahoo, Google and Overture are currently the most popular search engines on the Web in their respective categories [1].

3.1 The Test Suite

The test suite consisted of past examination questions administered to students at the Division of Information Studies at Nanyang Technological University. These questions covered a variety of subject areas taught at the graduate level by the Division, and may be broadly categorized into information technology-related (IT) topics and information/library services-related (ILS) topics. The former included areas such as

Internet and networking technologies, database management systems, information retrieval and human computer interaction, while the latter category included areas such as information sources, information organization, knowledge management and school media resource centers.

To reduce ambiguity in the experiments due to relevance judgments, only objective-type questions with known correct answers were used. This resulted in a total of 110 candidate questions from both categories. Of these, 31 questions were randomly selected to form the test suite of which 15 belonged to the ILS category with the remainder (16) falling into the IT category. Table 1 provides a sample of the questions selected.

Table 1

Sample questions from the test suite

List the skills in the Big Six Skills approach to problem solving.
List Ranganathan's laws of library science.
Identify three of the core professional values of librarianship.
What is the coverage of Books In Print?
Define query by example.
Which layers of the OSI model do brouters operate on?
What is a WIMP interface?
What is Z93.50?

3.2 Query Formulation

The next step involved formulating query expressions for each question. Two issues had to be addressed. The first concerned the query syntax for each of the two search engines. However this did not pose a major problem because both search engines simply required users to enter search terms without the need for operators. Consequently, the query expression used for each question consisted only of search terms and was identical for both search engines.

The second issue concerned the query expressions to use for the test suite. This is a potential source of bias as these expressions may vary between users in the actual terms used as well as in the number of terms used. To address this problem, two people familiar with Google and Overture were asked to independently formulate a query expression for each of the questions in the test suite. Disagreements were then resolved through discussion to arrive at a final "optimal" query expression.

3.3 Search Procedure

Once the expressions for the 31 questions were determined, each query was submitted to Google immediately followed by its submission to Overture. Searches on the entire test suite were completed on the same day. The rationale for this approach was to reduce the likelihood of changes to the search engine's indexes between search operations influencing the results of the study [14, 16].

For each query to a particular search engine, pages containing the first 20 results were saved into a file. While 20 items may appear arbitrary, the chosen number was not made without precedent with many studies adopting this figure (see for example [14, 17]). This

is further supported in a study of queries submitted to the Excite search engine conducted by Jansen, Spink and Bateman [18]. The authors found that most users (77%) viewed only the first 2 pages (10 results per page) of results.

3.4 Evaluation Procedure

When the search process was completed, the top 20 Web pages for each query were downloaded resulting in a total of 1240 documents to be evaluated. For better organization, one directory was created for each search engine to contain the results. Within each directory, the Web pages were stored in separate subdirectories named after question numbers. Directory names did not indicate the search engine used.

As discussed, the examination questions selected for this study had known correct answers. These were given to an evaluator who had to determine if a Web page for a particular query and search engine contained these answers. The evaluator had to read each Web page and indicate on a coding sheet its relevance according to the correct answer provided. Specifically, a Web page was considered *relevant* if it answered the question fully, *partially relevant* if it did not answer the question fully but contained a link to another Web page that did, and *irrelevant* if it did not answer the question and did not contain links to relevant Web pages.

The rationale for these criteria assumed that in an educational setting, students' have time constraints for completing their assignments which are then graded according to the quality of their answers. In cases where questions are objective in nature, grades would be awarded according to how closely an answer matches the actual answer. Consequently, a student searching the Web would likely consider only those documents that provided

complete answers or examine documents "near" the retrieved one. In the latter case, proximity is defined by the number of links traversed from a starting Web page and is fixed at one for this study. For example, for a question such as "List the skills in the Big Six Skills approach to problem solving" in Table 1, a student would probably not consider a document that described only one skill. He or she might examine other documents by clicking on available links but would soon retrieve other documents from the search results listings if under pressure to meet a deadline.

4. Results and Discussion

A two-way analysis of variance was conducted using question type (IT-type questions or ILS-type questions) and search engine type (Google or Overture) as the independent variables and precision as the dependent variable. Two definitions of relevance for computing precision were adopted. In the first (strict) case, only documents marked as "relevant" was considered relevant while in the second (lenient) case, documents marked as "relevant" or "partially relevant" were considered relevant.

Using the strict relevance criterion for precision, a significant main effect for search engine type was found, $F = 4.03$, $p < .05$. The mean precision values for Google and Overture were .40 and .30 respectively. This suggests that individuals using Google retrieve more relevant documents compared to individuals using Overture. The results thus strengthen the argument by Brin and Page [5] and others that traditional search engines perform better than PFP search engines in meeting the information needs of users. However, the results also mean that on average, 60% of documents returned by Google will be irrelevant documents while 70% of those returned by Overture will be

irrelevant. These numbers are high and suggest the need for further refinement of Google's ranking algorithm and Overture's relevancy guidelines.

A similar 2 (question type) X 2 (search engine) analysis of variance (ANOVA) was computed on the lenient relevance criterion for precision. This time however, the main effect for search engine type was not significant with the mean precision values for Google and Overture at .52 and .47 respectively. In other words, if users adopt the strategy of following promising links found in a retrieved document, Overture's performance will be comparable to Google's. However, given the time and effort needed to decide which links to follow and then to examine these documents, this approach is inefficient and users would be better off using a traditional search engine instead. In sum, the results indicate that Overture does not retrieve documents that users are looking for because a mismatch exists between the meanings assigned to the keywords by vendors and the meanings assigned by users. Stated differently, the results suggest that vendors value "low-relevance" keywords at the same (or misleading) rate as "high-relevance" keywords. Consequently, users have to look elsewhere to retrieve relevant documents from supposedly relevant documents.

Interestingly, closer analysis revealed a significant main effect for question type, $F = 3.98$, $p < .05$. For the lenient precision criterion, question type significantly influenced the number of relevant documents retrieved. The mean precision for IT-type questions was .55 while the mean precision for ILS-type questions was .44 suggesting that users would have a better chance of locating documents related to computers and technology than those related to the library sciences regardless of the type of search engine used. This means that ILS-type information needs should be supplemented with alternative

information sources. A possible reason for the lower means could be that individuals and groups in ILS fields may not have the expertise to develop Web sites and/or submit these to the major search engines. A further reason could be due to the non-commercial nature of library-related disciplines which translates to fewer documents being indexed in PFP search engines.

An analysis of document distribution was also conducted to determine if relevant documents were ranked higher by both search engines. Each search engine returned 20 documents and they were divided into four groups ranked according to position. The first group contained the first five documents (position 1-5) returned, the second group contained the next five documents (position 6-10) returned and so on. Table 2 shows the mean precision values for each group of five documents sorted by search engine type and relevance criterion.

Table 2

Means for search engine type and relevance criterion

Search Engine Type	Relevance Criterion	Subset			
		1	2	3	4
Google	Strict	.56	.40	.36	.40
Overture	Strict	.44	.32	.26	.28
Google	Lenient	.64	.50	.48	.48
Overture	Lenient	.56	.52	.40	.40

Note: Subset 1 = documents 1-5 returned; Subset 2 = documents 6-10 returned; Subset 3 = documents 11-15 returned; Subset 4 = documents 16-20 returned.

Two one-way ANOVAs were then conducted using precision (strict relevance criterion) as the dependent variable with position of documents returned (first five documents returned, next five returned etc.) by search engine type Google and Overture as independent variables in each of the ANOVA analysis respectively. Likewise, two one-way ANOVAs were conducted using precision (lenient relevance criterion) as the dependent variable with position of documents returned from Google and Overture as independent variables in each of the analyses. All four one-way ANOVAs yielded a significant main effect for position. Post-hoc comparisons were conducted with the Tukey's test for each significant *F* obtained. The pattern of results from the Tukey's procedure indicated that regardless of search engine type or precision criterion used, there was a statistically significant difference between the means of the first five documents returned and means of each block of subsequent documents returned. Further, the means of document positions of 6-10, 11-15 and 16-20 did not differ significantly from each other. Stated differently, the first five documents returned appear to be significantly more relevant compared to documents occupying lower ranked positions.

While the results suggest the soundness of the ranking algorithms used by both search engines, a final experiment was conducted to determine if there were any differences in document distribution between them. Due to number of documents retrieved and since many search engines (e.g. Google and Lycos) default to 10 documents per search results page, the retrieved documents were divided into two groups of 10 (first 10 documents retrieved and next 10 documents retrieved) with each group simulating one search results page. A 2 (search engine type) X 2 (document group) ANOVA with precision (strict relevance criterion) as the dependent variable yielded significant main effects for both

search engine type [$F = 12.06$, $p < .05$] and document group [$F = 12.06$, $p < .05$] but no significant interaction effect [$F = 0.027$, *ns*]. The means for Google and Overture were .43 and .33 respectively and the means for the first group of 10 documents and the next group of 10 were also .43 and .33 respectively. Google thus outperforms Overture in the ranking of relevant documents, with more relevant documents in Google being placed in the first group of 10 documents than for Overture. Once again, this suggests that while PFP search engines may be better for vendors because results can be influenced through paying for position within the search results, they may be biased against the needs of users who can be served better when position accurately reflects the relevance of retrieved documents.

A similar ANOVA conducted with the lenient relevance criterion for precision as the dependent variable yielded a significant main effect for document group [$F = 16.94$, $p < .05$] but search engine type only approached significance [$F = 3.88$, *ns*]. The mean for the first document group was .56 and the mean for the second was .44. This suggests that higher ranked documents appear to be more relevant than lower ranked documents regardless of search engine type. At first glance, this suggests that the ranking algorithms used by both search engines match users' expectations. However, the lenient relevance criterion requires users to examine more documents by following promising links within a purportedly relevant document. Taken together, it may thus be interpreted that for PFP search engines such as Overture to achieve mean precision values comparable to traditional search engines, users have to expend more effort by performing relevance judgments on retrieved documents.

5. Conclusion

The retrieval effectiveness of a search engine may be measured by precision and the distribution of relevant documents over the number of documents retrieved. The experiments conducted indicate that Google outperformed Overture in both aspects. Specifically, mean precision for Google was higher than for Overture meaning that more relevant documents were retrieved from the 20 returned. Further, Google appeared to rank documents more accurately so that more relevant documents were positioned near the top of the search results listings for Google than for Overture. Stated differently, these results suggest that Google users could find more documents with correct answers and that these documents could be retrieved faster because they were ranked near the beginning of the search results listings.

While the use of the lenient relevance criterion for precision reduced these disparities so that precision and distribution were comparable, a major drawback is that this criterion required more effort on the part of the user because more documents had to be examined by following links in a retrieved document. While users such as casual browsers may not find this too inconvenient, other users who require correct answers in the shortest amount of time might be better off with a traditional search engine.

In summary, this study lends support to the claim that PFP search engines produce biased results [5, 8] because it is the amount of money paid and not content characteristics that determines the relevancy of a document for a given query. In an academic setting, this especially holds true because information is often required on a subject or concept that is non-commercial in nature or is not affiliated with a particular brand or product. Consequently, PFP search engines may either return a small number of documents (low

recall) or many irrelevant documents (low precision) because the ultimate goal of a vendor is to sell products rather than provide free information with little hope of reaping any financial rewards.

It should be noted that the current study focuses on educational information needs and any conclusions should be restricted to this domain. In addition, while the study selected a major search engine from each category (traditional and PFP) according to popularity and/or size, the findings may not be applicable to other search engines. As such, this research is being expanded to include more search engines and more areas with both commercial and non-commercial interests as well. Examples include tourism, general news, shopping etc. It is expected that the results obtained will be better able to compare the performance of traditional search engines against their PFP counterparts.

References

[1] *Global Top 50 Web and Digital Media Properties*. Available at: http://www.jmm.com/xp/jmm/press/globalTop50WebProperties.xml.

[2] *Google Advertising Programs*. Available at: http://www.google.com/ads/.

[3] *Lycos Network Advertising Products*. Available at: http://adreporting.lycos.com/mediakit/lnap.html.

[4] D. Sullivan, Search engine sizes, *The Search Engine Report* (August 2001). Available at: http://www.searchenginewatch.com/reports/sizes.html.

[5] S. Brin and L. Page, The anatomy of a large-scale hypertextual web search engine, *Computer Networks and ISDN Systems* 30(1-7) (1998) 107-117.

[6] Overture – Create Search Listings. Available at: https://signup.overture.com/s/dtc/signup/.

[7] Overture – Advertiser's Guide to Listing with Overture. Available at: http://www.overture.com/d/about/advertisers/relevancy.jhtml

[8] P. Festa, GoTo gains amid Web pains, *CNET News.com* (2001). Available at: http://news.cnet.com/news/0-1005-200-5876770.html?tag=lh.

[9] M. Marchiori, The quest for correct information on the Web: hyper search engines. In: *Proceedings of the Sixth International World Wide Web Conference, Santa Clara, California* (1997) 265-276.

[10] D. Sullivan, Can portals resist the dark side? *The Search Engine Report* (June 2001). Available at: http://www.searchenginewatch.com/sereport/01/06-darkside.html.

[11] V. Kopytoff, Searching for profits, *San Francisco Chronicle* (June 18, 2001). Available at: http://www.sfgate.com/cgi-bin/article.cgi?file=/chronicle/archive/2001/06/18/BU107908.DTL.

[12] *Commercial Alert Files Complaint Against Search Engines for Deceptive Ads* (July 16, 2001). Available at: http://www.commercialalert.org/releases/searchenginerel.html.

[13] H. Chu and M. Rosenthal, Search engines for the World Wide Web: A comparative study and evaluation methodology. In: *Proceedings of the 59th ASIS Annual Meeting, Baltimore, Maryland* (1996) 127-135.

[14] H. Leighton and J. Srivastava, First 20 precision among World Wide Web search services (search engines), *Journal of the American Society for Information Science* 50(10) (1999) 870-881.

[15] M. Gordon and P. Pathak, Finding information on the World Wide Web: The retrieval effectiveness of search engines, *Information Processing and Management* 35(2) (1999) 141-180.

[16] J. Bar-Ilan, Evaluating the stability of the search tools Hotbot and Snap: A case study, *Online Information Review* 24(6) (2000) 439-449.

[17] W. Ding and G. Marchionini, A comparative study of Web search service performance. In: *Proceedings of the 59th ASIS Annual Meeting, Baltimore, Maryland* (1996) 136-142.

[18] B. Jansen, A. Spink, A, J. Bateman and T. Saracevic, Real life information retrieval: A study of user queries on the web, *SIGIR Forum* 32(1) (1998) 5-17.