Relevancy rankings : pay for performance search engines in the hot seat

Goh, Dion Hoe-Lian; Ang, Rebecca P.

2003

Goh, D. H. L., & Ang, R. P. H. (2003). Relevancy rankings: Pay for performance search engines in the hot seat. *Online Information Review*, 27(2), 87-93.

<https://hdl.handle.net/10356/91700>

<https://doi.org/10.1108/14684520310471699>

Goh, D.H., and Ang, R.P. (2003). Relevancy rankings: Pay for performance search engines in the hot seat. *Online Information Review*, 27(3), 87-93.

Relevancy Rankings – Pay for Performance Search Engines in the Hot Seat

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

Dion Goh is an assistant professor at the School of Communication and Information, Nanyang Technological University, Singapore

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

Rebecca Ang is an assistant professor at the National Institute of Education, Nanyang Technological University, Singapore

Relevancy Rankings – Pay for Performance Search Engines in the Hot Seat

Keywords

Information retrieval, Pay for performance, Search engines, Evaluation

Abstract

Pay for performance (PFP) search engines provide search services for documents on the Web but unlike “traditional” search engines, they rank documents not on content characteristics but according to the amount of money the owner of a Web site is willing to pay if a user visits the Web site through the search results pages. A study was conducted to compare the retrieval effectiveness of Overture (a PFP search engine) and Google (a traditional search engine) using a test suite of general knowledge questions. Forty-five queries based on a popular game show “Who Wants to be a Millionaire” were submitted to each of these search engines and the first 10 documents returned were analyzed using different relevancy criteria. Results indicated that Google outperformed Overture in terms of precision and number of queries that could be answered. Implications for this study are also discussed.

Introduction

Search engines have traditionally assisted users in filtering the glut of information on the World Wide Web through the concept of relevancy ranking in which search results are sorted according to algorithms that determine how closely a document matches a query. The criteria used by these ranking algorithms vary but are typically based on document

characteristics. These include number and frequency of matching terms, location of terms within the document, link structure and so on.

In recent years, several search engines have departed from the notion of ranking via document characteristics to more commercially oriented approaches. Known as pay for performance (PFP) search engines, these services rank retrieved Web pages in relation to the amount of money paid for desired search terms. To submit a site to a PFP search engine, the owner (or other stakeholders) bids on keywords that appropriately describe the Web site. This bid amount will be paid each time a user visits the site when it appears in the PFP search results listings. Typically, higher bids will mean a higher ranking for the site, and hence a greater likelihood that a user will visit that Web site when a search is executed. Overture, FindWhat and Sprinks are examples of popular PFP search engines.

The listings provided by PFP search engines are comparable to advertisements in that they serve the same role as other forms of advertising products offered by other search engines. Examples include targeted banner advertisements, “featured” links that appear at prominent locations of the search results page, and pop-up windows in which specific advertisements are displayed depending on the query terms entered.

The PFP Controversy

Unlike these advertising products however, PFP listings are somewhat controversial (Festa, 2001; Henshaw, 2001) because what appears as relevant Web pages to a user’s query are in reality, advertisements. While some PFP search engines do inform users of the nature of the search results, others do not, especially Web sites that contract with PFP search engines to provide search services for their visitors. Sullivan (2002) provides a listing of such Web sites (“search providers”) and the search engines contracted with.

The failure of some search providers and search engines to adequately inform users that paid listings are returned within search results has led to a complaint being filed to the US Federal Trade Commission (FTC) by Consumer Alert, a United States-based consumer group (Commercial Alert, 2001). 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.” In response to this complaint, the FTC issued a warning to these search engines, “outlining the need for clear and conspicuous disclosures of paid placement, and in some instances paid inclusion, so that businesses may avoid possible future commission action.” (Gallagher, 2002)

Rationale for Study

In an attempt to increase revenue and maintain (or attain) profitability, search engines are increasingly turning towards the PFP model of advertisement. However, do such services meet the information needs of users? A previous study by the authors (Goh and Ang, 2002) suggested otherwise. The retrieval effectiveness of Overture (a PFP search engine) and Google (a traditional search engine) was compared from an academic perspective. 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 indicated that Google outperformed Overture in both categories.

As the study focused on educational information needs, conclusions about search engine performance should be restricted to this domain. The present study thus seeks to extend this work to include questions covering a broader area. As in the previous study, Google (representing a traditional search engine) and Overture (representing a PFP search engine) were used. Forty-five general knowledge queries in different areas (e.g. sports, science, entertainment, history) were submitted to these services and the first 10 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 comparing the effectiveness of PFP search engines against their traditional counterparts.

Related Work

The study of search engine performance is an active area of research and the literature provides numerous examples. These include Ding and Marchionini (1996), Leighton and Srivastava (1999), and Gordon and Pathak (1999) to name a few. Each study offers a different perspective of querying on the Web and taken together, these studies present valuable information to researchers, search engine providers and users in the design, development and use of search engines.

To date however, studies comparing the retrieval effectiveness of PFP search engines against traditional search engines are rare, with most reports being anecdotal and sometimes contradictory. For example, Festa (2001) observes that in an attempt to be ranked highly for a given query, stakeholders of less popular Web sites would typically bid higher amounts for keywords in PFP search engines. Consequently, these sites would appear at or near the top of the search results pages and users are likely to get documents that are opposite of what they were looking for.

Similarly, Brin and Page (1998) argue that advertising funded search engines are biased towards advertisers and against the needs of users. Using an early PFP search engine named OpenText as an example, the authors contend that “this type of bias is much more insidious than advertising” because the search results listings do not clearly indicate if certain Web pages are ranked according to their document characteristics or according to the amount of money bid for keywords.

Sullivan (2001) 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.

To address the lack of research in PFP search engine performance, Goh and Ang (2002) conducted a study comparing the retrieval effectiveness of traditional search engines and PFP search engines using Google and Overture respectively. The test suite consisted of 31 queries extracted from past examination questions administered to students from a Master’s program in Information Studies, and included areas such as Internet and networking technologies, database management systems, information retrieval, human computer interaction, information organization, knowledge management and school media resource centers.

Each query was submitted to the two search engines and the first 20 documents returned were calculated for precision using two criteria for relevance. While several experiments

were conducted in the study, two important results are highlighted here. Firstly, the study revealed that mean precision for Google was significantly larger than Overture suggesting that traditional search engines perform better than PFP search engines in meeting the information needs of users. Secondly, results also indicated that Google outperformed Overture in the ranking of relevant documents, with more relevant documents in Google being placed higher in the search results listings than for Overture. This suggests that while PFP search engines may be better for stakeholders of Web sites because rankings 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.

Methodology

The present study seeks to replicate the earlier work but in a different domain, focusing on general knowledge questions. This is a sufficiently broad area that would appeal to a larger segment of users as compared to questions pertaining to graduate-level courses in Information Studies. As in the previous study, several experiments were conducted to compare the performance of a traditional search engine (Google) and a PFP search engine (Overture). 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.

The Test Suite

To introduce a sense of competition between the search engines in this study, the queries used were based on the television series “Who Wants to be a Millionaire”, a popular

game show originating from the United Kingdom in 1998 but has since spawned more than 100 country-specific versions from all over the world including the United States, India, Japan, Columbia, Australia, Russia and Singapore (Celador Productions, 2002). The questions asked cover a wide area and include subjects such as language, history, science, literature, television, movies, geography, animals and sports. A content analysis of the game can be found in Winzenburg (2000).

In the game, a contestant is asked a series of 15 multiple choice questions with increasing levels of difficulty, with each correct answer having a dollar value of between 100 to 1 million. If the contestant answers correctly, he/she moves to the next question while the game ends when a question is answered incorrectly. Questions are divided into two “safe levels” where a contestant is allowed to keep some of the prize money accumulated if a question is answered incorrectly. Specifically, if one of questions 1-5 is answered incorrectly, the contestant leaves with no money. If the contestant answers questions 6-10 incorrectly (the first “safe level”), he/she will leave with \$1000 in prize money. Finally, if the contestant answers questions 11-15 incorrectly (the second “safe level”), the prize money won will be \$32000.

A total of 390 questions worth differing dollar amounts were obtained from various “Millionaire” Web sites around the world containing actual quiz questions. These were divided into 3 categories representing questions 1-5 (“unsafe level”), the first “safe level” and the second “safe level”. Fifteen questions were then randomly selected from each category for a total of 45 questions. Table 1 provides a sample of the questions selected and their corresponding category.

Table I. Sample questions from the test suite

Question	Category
Who designed the Eiffel Tower?	Unsafe
Who is the lead singer in the group Eurythmics?	Unsafe
Sitting Bull was a chief of which native American people?	First
What is the currency of the kingdom of Tonga?	First
What disease is also known as “lymphadenoma”?	Second
Where was Freddie Mercury born?	Second

Query Formulation, Search Procedures

Query formulation followed the same procedures as in the previous study to maintain consistency. The goal of this phase was to develop queries that could be used by both search engines as well as to reduce biasness in their formulation among different users. Two people familiar with searching using Google and Overture independently formulated a query for each of the questions in the test suite. These query expressions consisted only of search terms and were identical for both search engines. Disagreements were then resolved through discussion to arrive at a final query expression.

Queries were then submitted to Google and Overture in succession and the entire test suite was completed in one session on the same day. The rationale for this approach was to reduce the likelihood of changes to the search engine’s indexes influencing the results of the study (Bar-Ilan, 2000). In the present study, the top 10 documents returned by the search engines were used for analysis. Thus, pages containing the first 10 results from each query to a particular search engine were saved into a file. Using only the top 10

pages in search engine experiments is not uncommon (see for example Chu and Rosenthal (1996)) and is reflective of the search patterns of many Web users (see for example Spink and Xu (2000); Jansen, Spink and Saracevic (2000)). Finally, in the previous study, the authors (Goh and Ang, 2002) found a significant difference in precision between the first 10 documents returned and the next documents in both Google and Overture, suggesting that relevant documents are typically located in the first 10 pages of the search results listings. This finding will be elaborated in the Results section.

Evaluation Procedure

Using the saved search results pages, the top 10 Web pages were downloaded resulting in a total of 900 documents to be evaluated. These documents were then organized into separate directories by question number, and the resulting 90 directories (45 for each search engine) were further organized by search engine type within 2 directories. Directory names did not indicate the search engine used to prevent any biasness towards or against a particular search engine.

All questions had a known correct answer and were recorded in an answer key. This was provided to an evaluator who had to determine if a Web page for a particular query and search engine correctly answered a question. As in the previous study, the evaluator had to read each Web page and indicate on a coding sheet its *relevance type* according to the following criteria:

- *relevant* – the page answered the question fully
- *partially relevant* – the page did not answer the question fully but contained a link to another Web page that did

- *irrelevant* – the page did not answer the question and did not contain links to relevant Web pages

The rationale for these criteria assumes that users want to meet their information needs as quickly as possible. In the simplest case, a user would consider a Web page relevant if the answer to his/her query can be found directly in it. For Web pages that do not provide an answer however, the user might consider visiting other pages “near” the one currently being viewed. Here, proximity is defined by the number of links traversed from a starting Web page and is set at one for this study. In other words, it is assumed that if a user does not successfully obtain a correct answer from a document or from other documents that are one link away, he/she will examine other documents in the search results listings.

Results and Discussion

A two-way analysis of variance was conducted using categories of questions (unsafe level, first level or second level 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 *strict relevance* case, only documents marked as “relevant” were considered relevant while in the *lenient relevance* case, documents marked as “relevant” or “partially relevant” were considered relevant.

A significant main effect for search engine type was found using the strict relevance criterion for precision as the dependent variable, $F = 10.04$, $p < .05$. The mean precision values for Google and Overture were .71 and .50 respectively, indicating that Google users retrieve more relevant documents compared to Overture users. This strengthens the argument by the authors, Brin and Page (1998) and others that traditional search engines

perform better than PFP search engines in meeting the information needs of users. Conversely, the results also mean that on average, 30% of documents returned by Google will be irrelevant while 50% of those returned by Overture will be irrelevant. These figures are lower compared to the ones obtained in the earlier study by Goh and Ang (2002) – 60% and 70% respectively. The better scores are perhaps due to the fact that general knowledge type questions are more easily available on the Web than academic type questions because of the broader appeal of the former. Nevertheless, a .50 precision value for Overture indicates a need for further refinement of Overture's relevancy guidelines (Overture, 2002).

A similar 3 (categories of questions) X 2 (search engine) analysis of variance was computed on the lenient relevance criterion for precision. Likewise, a significant main effect for search engine type was found, $F = 10.55$, $p < .05$. The mean precision values for Google and Overture were .76 and .56 respectively. This is in contrast to the previous study in which no significant main effect for search engine type was found. Taken together, the results appear to suggest that Google's performance is superior to Overture's especially in general knowledge type questions. The results also reinforce that view that both search engines provide better coverage of general knowledge type questions as compared to academic type questions. This is demonstrated by the relatively small increase in precision values of 7% (Google) and 12% (Overture) in the current study versus a 30% (Google) and 56% (Overture) increase in the previous study.

An analysis of document distribution was also conducted to determine if relevant documents were ranked higher by both search engines. In the previous study (Goh and Ang, 2002), a significant difference in precision between the first 10 documents returned

and the next 10 documents in both Google and Overture was found, suggesting that relevant documents are typically located in the first 10 pages of the search results listings. For Google, the first 10 documents returned ($M = .38$) were ranked significantly higher than the next 10 documents ($M = .27$) returned, $t = 3.05$, $p < .05$. Likewise, the first 10 documents returned ($M = .54$) were ranked significantly higher than the next 10 documents ($M = .40$) returned, $t = 3.50$, $p < .05$ for Overture. This provided empirical support for using only the top 10 documents for analysis in this study. In the present study, each search engine returned 10 documents and they were divided into two groups ranked according to position. The first group contained the first five documents (position 1-5) returned while the second group contained the next five documents (position 6-10) returned. Table 2 shows the mean precision values for each group of five documents sorted by search engine type and relevance criterion.

Table II. Means for search engine type and relevance criterion

Search	Engine	Relevance	Subset	
Type		Criterion	1	2
Google		Strict	.78	.64
Overture		Strict	.54	.42
Google		Lenient	.90	.72
Overture		Lenient	.62	.48

Note: Subset 1 = documents 1-5 returned; Subset 2 = documents 6-10 returned.

Two T-tests were then conducted using precision (strict relevance criterion) as the dependent variable with position of documents returned (first five documents returned, and next five returned) from Google [$t = 2.56$, $p < .05$] and Overture [$t = 3.80$, $p < .05$] as

independent variables in each of the t-test analyses respectively. Likewise, two independent t-tests were conducted using precision (lenient relevance criterion) as the dependent variable with position of documents returned from Google [$t = 4.81, p < .05$] and Overture [$t = 4.95, p < .05$] as independent variables in each of the analyses. All four t-tests yielded a statistical significance for position (see the means presented in Table 2). The pattern of results indicate that there was a statistically significant difference between the means of the first five documents returned and the means of the next five documents returned. Stated differently, the first five documents returned appear to be significantly more relevant compared to the next five documents occupying lower ranked positions, suggesting the soundness of the ranking algorithms used by both search engines.

Finally, two chi-square tests were conducted (for strict and lenient relevance criteria) to test if there were differences between Google and Overture on queries resulting in zero documents returned. For the strict relevance criterion, Google was found to have significantly fewer zero hits compared with Overture ($\chi^2 = 8.78, p < .01$). Similarly, for the lenient relevance criterion, Google was also found to have significantly fewer zero hits compared with Overture ($\chi^2 = 7.59, p < .01$). The results thus suggest a broader coverage of topics in Google than in Overture, reflected in part by the index sizes of the two search engines. Stated differently, Google users will be more likely to satisfy their information needs than Overture users because fewer of their queries will result in zero documents returned.

Conclusion

With the growth in the number of PFP search engines on the Web and their alliances with portals/directories and other search providers (Sullivan, 2002), searchers will increasingly

use these services directly or indirectly. The present study attempts to determine if the information needs of users can be met through PFP search engines by comparing the retrieval performance of Overture and Google.

The test suite consisted of 45 general knowledge type questions based on a popular game show “Who Wants to be a Millionaire”, and analysis was based on the top 10 documents from each query returned by both search engines. Retrieval performance was measured by precision using 2 definitions of relevance (strict and lenient) and the number of queries that resulted in zero hits. The experiments conducted indicate that Google outperformed Overture in both aspects. Specifically, mean precision for Google (.71 for strict relevance; .76 for lenient relevance) was higher than for Overture (.50 strict relevance; .56 lenient relevance) meaning that more relevant documents were retrieved from the 10 returned. Further, Google appeared to provide a broader coverage of topics than Overture as reflected by a significant difference in the number of queries with zero hits. Put together, these results suggest that Google users can have more of their queries answered and find more documents with correct answers than Overture users.

In summary, this study confirms the conclusions drawn in the previous study by Goh and Ang (2002) and others that PFP search engines produce biased results because it is the sum of money paid for keywords and not document characteristics that determines the relevancy of a document for a given query. For general knowledge type questions, this observation holds true because the information need is often related to a topic or concept that is non-commercial in nature or is not associated with a specific brand or product. As a result, PFP search engines may either return a small number of documents (low recall) or many irrelevant documents (low precision) to a query because few stakeholders of

Web sites would be willing to provide free information with little hope of reaping any financial rewards.

The present work is part of a series of studies comparing the performance of PFP search against their traditional counterparts. A future research direction in this area will investigate retrieval performance with more search engines using queries where information needs relate to products and services. It is expected that with the results obtained in this future study in combination with the existing work conducted by the authors, users will be better able to make informed decisions regarding the use of PFP search engines.

References

- Bar-Ilan, J. (2000), "Evaluating the stability of the search tools Hotbot and Snap: A case study", *Online Information Review*, Vol. 24 No. 6, pp. 439-49.
- Brin, S. and Page, L. (1998), "The anatomy of a large-scale hypertextual web search engine", *Computer Networks and ISDN Systems*, Vol. 30 No. 1-7, pp. 107-17.
- Celador Productions (2002), *Who Wants To Be A Millionaire*, available at:
<http://www.millionaireonline.tv/enter.html>
- Chu, H. and Rosenthal, M. (1996), "Search engines for the World Wide Web: A comparative study and evaluation methodology", *Proceedings of the 59th ASIS Annual Meeting, October 21-24, Baltimore, Maryland*, pp. 127-35.
- Commercial Alert (2001), *Commercial Alert Files Complaint Against Search Engines for Deceptive Ads*, available at:
http://www.commercialalert.org/index.php?category_id=1&subcategory_id=24&article_id=33.

- Ding, W. and Marchionini, G. (1996), "A comparative study of Web search service performance", *Proceedings of the 59th ASIS Annual Meeting, October 21-24, Baltimore, Maryland*, pp. 136-42.
- Festa, P. (2001), "GoTo gains amid Web pains", *CNET News.com*, available at:
<http://news.cnet.com/news/0-1005-200-5876770.html?tag=lh>.
- Gallagher, D. (2002), "U.S. warns Web sites to label sponsorships", *The New York Times*,
 available at: <http://www.nytimes.com/2002/07/02/technology/02SEAR.html>
- Goh, D. and Ang, R. (2002), "Are pay for performance search engines relevant?",
Journal of Information Science, Vol. 28 No. 5, pp. 349-55.
- Gordon, M. and Pathak, P. (1999), "Finding information on the World Wide Web: The retrieval effectiveness of search engines", *Information Processing and Management*, Vol. 35 No. 2, pp. 141-80.
- Henshaw, R. (2001), "What next for Internet journals? Implications of the trend towards paid placement in search engines", *First Monday*, Vol. 6 No. 9, available at:
http://firstmonday.org/issues/issue6_9/henshaw/index.html
- Jansen, B., Spink, A. and Saracevic, T. (2000), "Real life, real users and real needs: A study and analysis of users queries on the Web", *Information Processing and Management*, Vol. 36 No. 2, pp. 207-27.
- Leighton, H. and Srivastava, J. (1999), "First 20 precision among World Wide Web search services (search engines)", *Journal of the American Society for Information Science*, Vol. 50 No. 10, pp. 870-81.
- Overture (2002), *Listing Guidelines*, available at:
<http://www.overture.com/d/about/advertisers/relevancy.jhtml>

- Spink, A. and Xu, J. (2000), "Selected results from a large study of Web searching: The Excite study", *Information Research*, Vol. 6 No. 1, available at:
<http://informationr.net/ir/6-1/paper90.html>.
- Sullivan, D. (2001), "Can portals resist the dark side?", *The Search Engine Report*,
available at: <http://www.searchenginewatch.com/sereport/01/06-darkside.html>.
- Sullivan, D. (2002), "Who powers whom? Search providers chart", *The Search Engine Report*, available at: <http://www.searchenginewatch.com/reports/alliances.html>
- Winzenburg, S. (2002), *A Content Analysis of "Who Wants To Be A Millionaire"*,
available at: http://faculty.gvc.edu/swinzenburg/millionaire_study.html