Searching the Wikipedia with Public Online Search Engines

Miro Lehtonen

Department of Computer Science
P.O. Box 68 (Gustaf Hällströmin katu 2b)
FI–00014 University of Helsinki
Finland
Miro.Lehtonen@cs.helsinki.fi

Abstract. Commercial search engines were put to a test as we searched the online Wikipedia which is a newer version of the INEX 2010 document collection. Although the INEX 2010 ad hoc search tasks and the search features of the public search engines are not 100% compatible, we were able to compare and evaluate the search results of online search engines with INEX 2010 topics, assessments, and metrics. Considering the first page of results, we cannot see a big difference between the performance of the best academic search engines and the best commercial ones. Of the public search engines, Google and Yahoo perform marginally better than Bing and significantly better than the on-site Wikipedia search.

1 Introduction

Public online search engines have not been seen in the official INEX evaluations in the past despite the highly competitive performance that they offer to web users. Whether they are competitive also when measured in such a standard IR evaluation has not been reported before, to the best of my knowledge. The purpose of this experiment was to find out how well some of the most popular search engines fare with the academic search engines and with each other. The search engines of choice are Google, Bing, Yahoo!, and the default Wikipedia search which specialises in searching the online version of the Wikipedia[1].

The comparison is not completely fair for a number of reasons. For one thing, the Wikipedia articles evolve constantly. Not only is the online Wikipedia different from the INEX version, but each online search engine indexes a slightly different version of the document collection, as well. For another thing, none of the search engines return 1500 results per query. It is possible to set a limit on the results per page, but ultimately, the number of retrieved results depends on the query. Sometimes the search engines do not return any results, which can be considered more user-friendly than returning nearly 1500 non-relevant results. User satisfaction is however not taken into account by the metrics of INEX. For example, novelty and diversity [3] are not rewarded; nor is wasted user

[1] This experiment has not been endorsed by any of the mentioned search engines.

S. Geva et al. (Eds.): INEX 2010, LNCS 6932, pp. 82–88, 2011.

effort penalized. Nevertheless, the results should be indicative of the true performance, given the variety of different metrics and appropriate interpretation of the results.

It is commonly known that web search engines rely on information outside the Wikipedia, including incoming links from other web pages as well as click data specific to each query and search engine. This might give them an advantage over the academic search engines that only rely on the INEX version of the Wikipedia. However, as the Web is available to all the INEX participants, we cannot assume that the presumed advantage is unfair.

The analysis of the results shows that the best academic search engines are just as competitive as the best commercial search engines when all the circumstances are taken into account, e.g. we only search the Wikipedia. Google and Yahoo are highly competitive when looking at the first page of results.

2 Related Work

Commercial online search engines have been compared in the past in various ways and various metrics. Measuring the popularity and profitability is without a doubt an interesting way to rank the search businesses by their success. A more technical aspect that is often measured is the size of the index or web coverage [5]. However, there is no search business without a search engine with an effective search algorithm. Previously, not many Cranfield-style evaluations of this scale on commercial search engines have ever been published. For example, Tümer et al. compared Google, Yahoo, MSN, and Hakia with 10 queries on the live WWW [6], whereas Bar-Ilan et al. compared the result lists of Google, Yahoo, and Teoma and measured overlap and statistical correlation between the lists [2].

A recent study worth noting was conducted by Ganjisaffar et al. [4] who pooled the top 10 results of Google, Yahoo, and Live search on the domain of *en.wikipedia.org*. In their experiment, seven assessors labelled a total of 240 queries resulting in a finding that Live search outperforms both Google and Yahoo.

3 Running Online Search Engines

The tests described in this section were conducted in August 2010. Although the results for individual queries may change over time as the search engines index and re-index updated pages, the changes should not affect the overall performance or the mutual rankings of the search engines.

All the search engines were used in a uniform fashion. One HTTP request was made for each INEX 2010 topic and for each search engine, after which the server response — the first page of results — was dumped into a file for further analysis and processing. The maximum number of results on the first page naturally depends on the search engine. The URLs for the HTTP requests came from entering the title field of the topic in the search box as written in the topic file. Quotation marks, plus and minus signs were all included as such.

3.1 Google

Of all the different Google search sites, the one that is not specific to any country was chosen[2] although the rankings might be stable across different country versions when searching a single site. The maximum number of results per page was set to 100.

Most of the results that Google returns also exist in the INEX version of the Wikipedia. The retrieved pages that are not found in the INEX collection are either new articles or combinations of old ones with a new article ID.

3.2 Bing

Bing seems to assume that different rankings are called for in different countries — even when searching a single site such as the English Wikipedia. Therefore, choosing a country version for this experiment makes a real difference. American bing[3] was chosen for except the assumption that, as one of the most frequently used country versions of Bing, it should be up-to-date and the rankings should be rather stable. The maximum number of results on the first page was set to 50 which is the biggest number allowed.

Unlike Google, Bing retrieves a fair amount of pages that are not included in the INEX collection although the content is. For example, one of the pages that Bing returns has the title of "Marilyn Munroe" and the URL

`http://en.wikipedia.org/wiki/Marilyn_Munroe`.

The page redirects to a page correctly titled "Marilyn Monroe" which does exist in the INEX collection and which is retrieved by Google as a link to

`http://en.wikipedia.org/wiki/Marilyn_Monroe`.

The contents of these two pages are equivalent, but, if the page is relevant to a query, only Google scores whereas Bing hits a missing page. None of the search engines retrieve both pages because they try to avoid including duplicate pages in the results. This might be a case where Bing is being unfairly penalized for retrieving "wrong" pages where the content is relevant but the title is misspelled.

3.3 Yahoo!

Yahoo[4] returns a maximum of 100 results on the first page with a note "Powered by Bing". The exact reliance on Bing is somewhat unclear but one of the prolems is the same: Yahoo too retrieves a number of redirected pages which are not included in the INEX version of Wikipedia.

3.4 Wikisearch

The on site Wikipedia search engine[5] is the only search engine in this experiment that returns focused results — in addition to generating snippets for each

[2] www.google.com/ncr
[3] www.bing.com/search?setmkt=en-US&q=...
[4] search.yahoo.com
[5] http://en.wikipedia.org/w/index.php?title=Special:Search&search=...

article. While other search engines retrieve whole articles from the Wikipedia, the Wikipedia search engine also suggests which section might be relevant to the query. However, this feature of Wikisearch was ignored because the anchor of the section under focus cannot be converted into an entry point without opening the INEX version of the article, and even then the conversion is not completely reliable.

Wikisearch allows a total of 500 results to be shown on the first page. Like Google, the on site search does not return any pages that redirect further, so most of the retrieved articles also exist in the INEX version of the Wikipedia.

4 From Result Pages to Run Submissions

Run submissions were created for two different tasks: Restricted Focused (RF) and Restricted Relevant in Context (RRIC). Processing the first page of results began the same way for both tasks. First, the article titles were collected by scanning the result page and extracting the title from the URL. Second, the corresponding pages were found in the INEX collection by matching the titles of the online article to the titles in the INEX 2009 version of the Wikipedia. Because the actual online article was not accessed, the reason for not finding a matching article in the INEX collection was not analysed. Once we had a ranked list of articles for each topic, we could create a task-specific run submission.

4.1 Restricted Focused

The ranked list of articles which was specific to each search engine did not contain any focused results. Therefore, the focus had to be artificially added to the result list. It had to be a blind process because the document and element scores of the search engines were not accessible and because we wanted to eliminate the effect of all external factors. A simple but heuristic way to meet the task requirement of 1000 characters per topic was to pick the top two articles from the list and return a passage consisting of the first 500 characters of each.

4.2 Restricted Relevant in Context

Creating a run submission for the RRIC task was straightforward. All of the articles on the first page of results were included. Restricting the results to 500 characters per article was a task requirement. Because none of the search engines provide ways to define such a restriction, a simple heuristic had to be defined for all of them. Assuming that the beginning of the article would be the best entry point, the first 500 characters of the article would be a good guess on the restricted passage. Retrieving 500 characters from the beginning of the article was also simple to implement.

The Wikipedia search is the only search engine where the first 500 characters are not always part of the result retrieved by the online search engine because Wikipedia search sometimes focuses the results to certain sections. In those cases, the real Wikipedia might get better scores than it gets in this experiment.

4.3 Summary

All the search engines are good at removing duplicate pages from the results so that the same content is not retrieved multiple times although it may exist under several different URLs. How many pages each search engine retrieved that also exist in the INEX Wikipedia collection is summarised in Table 1.

Table 1. Summary of submitted results for the 107 topics of INEX 2010

Search engine	First page	Total RF	Total RRIC	Max RRIC
Google	100	212	7,353	10,700
Bing	50	208	1,156	5,350
Yahoo!	100	209	5,893	10,700
Wikisearch	500	202	28,770	53,500

There were a total of 107 topics in 2010, so the maximum number of submitted results for the RF task would be 214 and for the RRIC task 160,500 (1500 results per query), given the chosen heuristics. The number of results that was actually retrieved is bigger than the number of submitted results because of new pages and redirecting pages.

5 Results

The evaluation for the runs submitted for the RF task is shown in Table 2. Although Google seems to retrieve relevant articles with the highest precision, Yahoo has the highest *character precision*, retrieving the highest ratio of relevant content to non-relevant content. However, limiting the results to 500 characters per article was merely an artificial post-search procedure to satisfy the task requirements, and therefore, the search engines should not be rewarded or penalised for it. As this comparison only considers the top two results for each query, it is fair to compare how many relevant articles the search engines ranked in the top two ranks of the result list. Google tops the chart as the only search engine that returns at least two results for each of the 52 queries included in the evaluation. Google also has the highest total number of relevant articles found (79) and the highest precision (75.96%).

Table 2. Evaluation of the top two results for the 52 topics submitted for the restricted focused task

Search engine	articles	relevant	art_prec	char_prec	iP0.01
Google	104	**79**	0.7596	0.3276	0.1040
Yahoo	102	**77**	0.7404	0.3435	0.1186
Bing	102	**75**	0.7212	0.3354	0.1062
Wiki	100	**68**	0.6538	0.2670	0.0713

Comparing the public search engines with the best academic ones, we note that Yahoo (0.3435) ranks the 3rd in the Restricted Focused task where the official measure was character precision [1]. Bing (0.3354) and Google (0.3276) are not far behind whereas Wikisearch has the lowest score — obviously due to the largest number of results per query.

Whether there is any significant difference between the results is tested in Table 3. According to the topic-wise t-test on article precision, both Google and Yahoo perform significantly better than the default Wikipedia search but the differences between other search engines are not statistically significant.

Table 3. P-values of the t-test (one-tailed)

T-test	Yahoo	Bing	Wiki
Google	0.2424	0.1260	**0.0074**
Yahoo		0.2989	**0.0138**
Bing			0.0818

The RRIC runs consist of the first page of results for each query before restricting the submission to the first 500 characters of the article. The performance of the official RRIC runs are shown in Table 4. The number of results returned on the first page varies a lot, which makes a direct comparison of absolute precision or recall unjustified. As we are interested in article precision at fairly low ranks, we order the runs according to interpolated precision at 0.01 which has been considered a rather stable measure in the past INEX evaluations.

Table 4. Evaluation of runs submitted for the restricted relevant in context task

Search engine	articles (avg)	relevant	art_prec	char_prec	iP0.01
Yahoo	2946 (56.7)	1075	0.3463	0.1290	**0.2848**
Google	3537 (68.0)	1305	0.3629	0.1264	**0.2658**
Wiki	13688 (263.2)	2344	0.1819	0.0463	**0.1995**
Bing	549 (10.6)	301	0.5363	0.0115	**0.1975**

As a side note, Bing and Yahoo results are slightly compromised because of the mismatch between titles returned and titles in the INEX collection. Moreover, all the search engines did slightly better live than in this experiment as they retrieved relevant articles which were more recent than the INEX 2009 document collection.

6 Conclusion

Four public online search engines were tested when searching the Wikipedia with the INEX 2010 topics. Google, Bing, Yahoo, and the on site search of Wikipedia all retrieve relevant articles from the Wikipedia with varying success.

What we learn from this experiment is that searching the Wikipedia with Google or Yahoo is worth a try when the Wikipedia search does not find any relevant articles. However, the only search engine that returned focused results was the Wikipedia search — a feature that will hopefully be appreciated by future users.

References

1. Arvola, P., Geva, S., Kamps, J., Schenkel, R., Trotman, A., Vainio, J.: Overview of the inex 2010 ad hoc track. In: INEX 2010. LNCS, vol. 6932, pp. 1–32. Springer, Heidelberg (2011)
2. Bar-Ilan, J., Mat-Hassan, M., Levene, M.: Methods for comparing rankings of search engine results. Comput. Netw. 50, 1448–1463 (2006)
3. Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, pp. 659–666. ACM, New York (2008)
4. Ganjisaffar, Y., Javanmardi, S., Lopes, C.: Leveraging crowdsourcing heuristics to improve search in wikipedia. In: Proceedings of the 5th International Symposium on Wikis and Open Collaboration, WikiSym 2009, pp. 27:1–27:2. ACM, New York (2009)
5. Kim, Y.S., Kang, B.H., Compton, P., Motoda, H.: Search engine retrieval of changing information. In: Proceedings of the 16th International Conference on World Wide Web, WWW 2007, pp. 1195–1196. ACM, New York (2007)
6. Tumer, D., Shah, M.A., Bitirim, Y.: An empirical evaluation on semantic search performance of keyword-based and semantic search engines: Google, yahoo, msn and hakia. In: Proceedings of the 2009 Fourth International Conference on Internet Monitoring and Protection, pp. 51–55. IEEE Computer Society, Washington, DC, USA (2009)