Ranking of Wikipedia articles on search engines for searches about its own articles

Seminar Task for Internet Search Techniques and Business Intelligence class

by Jure Čuhalev <jure.cuhalev at guest.arnes.si>

Friday, October 13, 2006

Introduction

During summer 2006 a lot people noticed that Wikipedia[1] is starting to rank very high for their searches on search engines, especially Google. In August 2006, Nicholas G. Carr wrote a series of blog posts (Carr, 2006a and 2006b) in which he began to explore possibilities of centralization of information in Wikipedia and noted that with time the importance of Wikipedia as an information source will only grow.

In September 2006, Steve Rubel published his study on Wikipedia's Impact on Top 100 U.S. Brands (Rubel, 2006) where he showed that many brands Wikipedia articles rank very high for Google searchs, very often in top 10 results. He highlights the fact that having a good entry in Wikipedia about brand is becoming increasingly more important for brand owner.

Most, if not all, of the observations around this topic deal with subjective experience or with limited research into different fields. Additionally there is a lot of research done on Wikipedia itself but since this phenomenon is relatively new, there has not been more extensive research into it.

In this seminar paper I will try to more systematically research the question of *how do Wikipedia articles rank for general search queries that have answer in Wikipedia*. To determine this I will analyze Wikipedia rank in top 10 search results for a sample of article titles from Wikipedia in three different search engines. Additionally I will try to determine if above observations are also valid in more general context.

[1] http://www.wikipedia.org - Wikipedia is a free online encyclopedia that is written by volunteers from all over the world. Contributions are open to everyone.

Methodology

As of time of the writing, Wikipedia consists of more than 1.4 million articles. Wikipedia also support *redirects* where one article title redirects to a different article. This, for example, can be used for topics that can be spelled in different ways or abbreviations. When user enters such redirect page Wikipedia does not redirect actual URL to an actual article but presents the linked articles and only displays a note on top of the page that this is a redirect. This means that search engines see those URL's as a separate page of its own. There are ~1.3 million redirects. This makes the total number of article titles worth considering ~2.7 million. Full article title list together with redirects is created on daily basis and is available for download[2].

From this article list, 1000 article titles were selected using simple random sampling method (Lehtonen, 2004 and Vehovar, 2001) and trying to achieve at least 95% confidence level at 5% standard error. For selection of articles a random number generator at random.org was used. After the articles were selected the list was manually filtered for titles with obscure Unicode characters leaving 992 article titles. This was done because of limitations in software used for analysis.

Each of the titles was converted into a search query by changing underscore into a space (e.g. Antonio,_The_Prior_of_Crato into Antonio, The Prior of Crato). No additional changes to titles were done.

Each search query was then submitted to main English versions of three search engines: Google[3], Yahoo[4] and MSN[5] using their API interfaces. Top 10 results from each query were then stored into a database and further analyzed. Tools used were Python and Django framework for Python, SPSS and Excel.

Additional definitions used were:

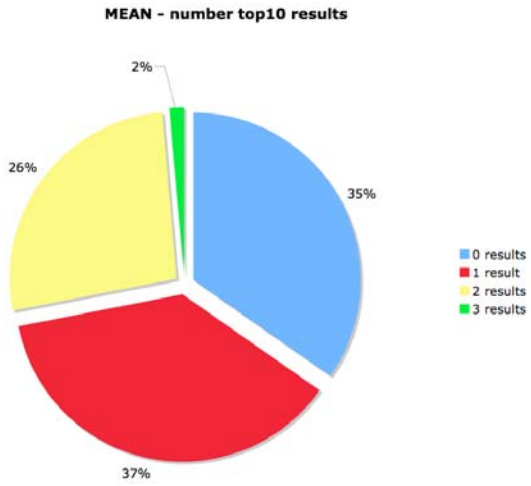
- Wikipedia link is any URL that contains string *wikipedia.org/wiki/*.
- Exact Wikipedia page is exact URL of Wikipedia title as it would be accessed via http://en.wikipedia.org/wiki/[page title].

Analysis

In analysis I will focus on ranks of Wikipedia articles and also try to highlight the differences between different search engines. Only more interesting results are presented below with full statistics available as Appendix B.

[2] http://download.wikimedia.org/ - Wikipedia database dumps
[3] http://www.google.com
[4] http://www.yahoo.com
[5] http://www.msn.com

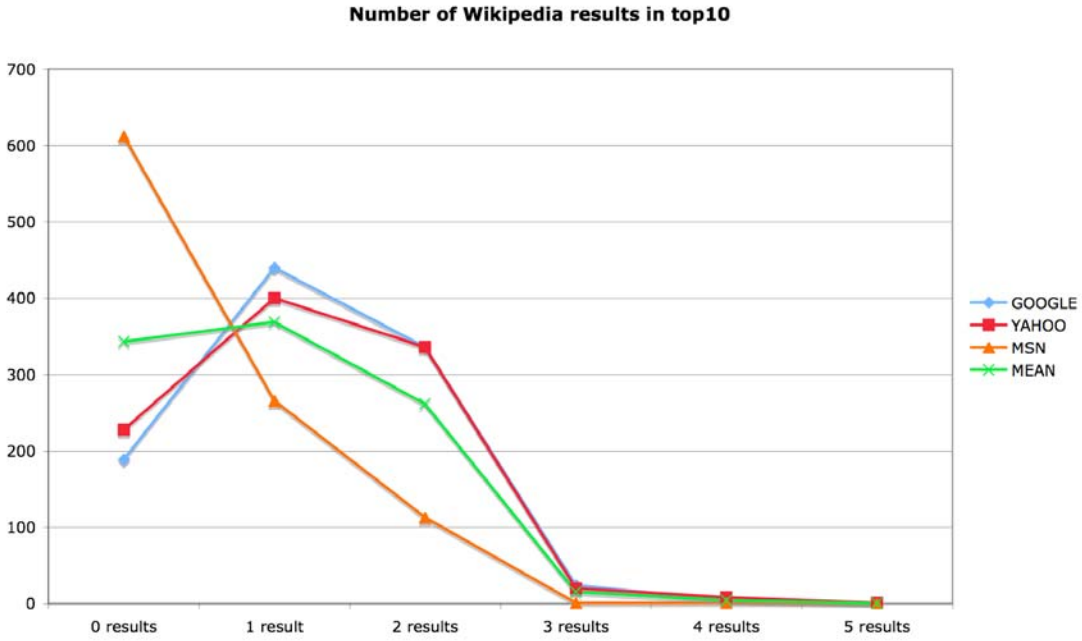
Number of results



For search terms, in average only 35% of queries did not have Wikipedia link in top 10 results, 37% of them had only 1 result, 26% 2 results and 2% had 3 results or more.

For specific search engines these numbers were:

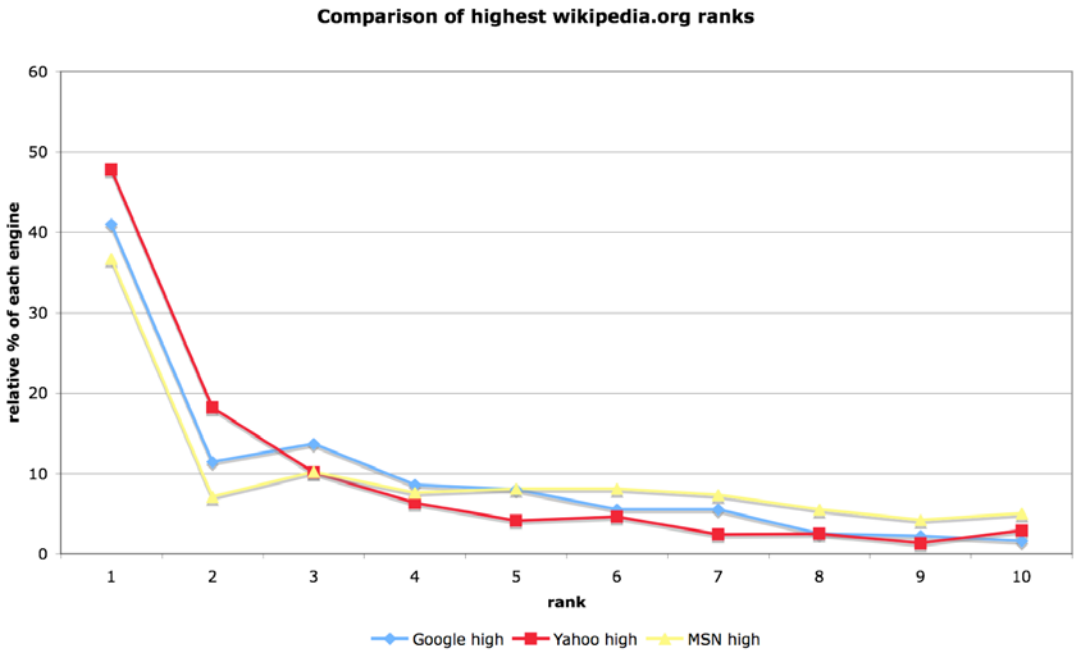
Number of results	Google (%)	Yahoo (%)	MSN (%)
0	*19.0*	*23.0*	*61.6*
1	44.3	40.3	26.8
2	33.8	33.8	11.4
3	2.4	2	0.1
4 or more	0.4	0.9	0.1



Observing these numbers regarding the specific search engines we can see that MSN search has noticeably lower number of Wikipedia pages in top 10 search results. Both Google and Yahoo prefer to list only 1 or 2 results with more being very uncommon.

Rank analysis

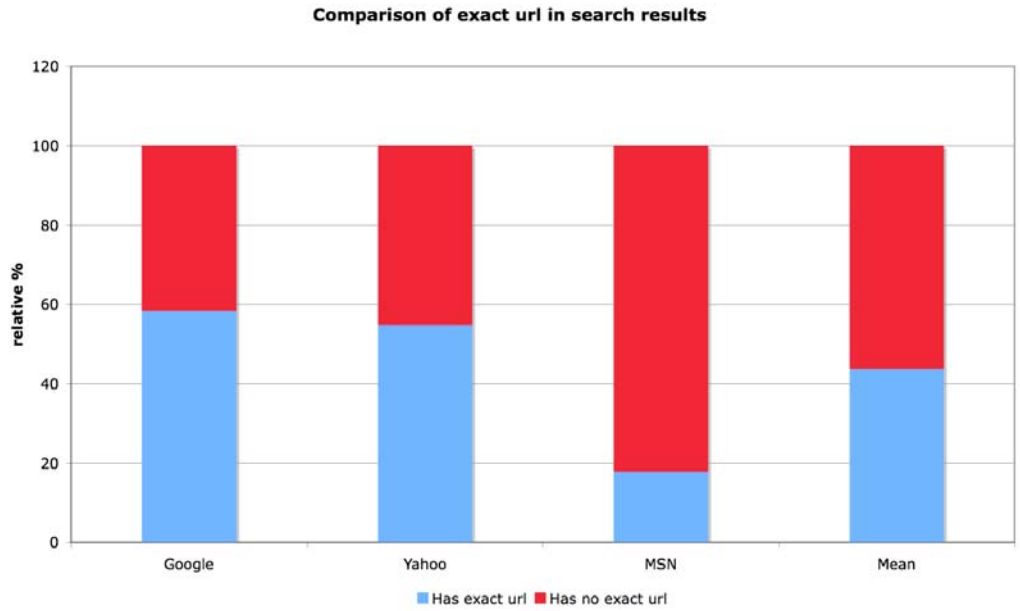
All results in this section, unless otherwise noted, are done only on cases where there was a Wikipedia article in top 10 results.



Observing trend lines we can see that Wikipedia ranks very highly on top 10. It is in top 3 results 66.2% of time for Google, 76.1% for Yahoo and 54.1% for MSN. We can see slight preference of MSN to not rank Wikipedia as the top result and Yahoo preferring it the most.

Which Wikipedia page

Another issue worth analyzing is if the article itself is considered to be best answer to search query by search engine.



We can see that search engine link to the exact article in average in only 43.7% (Google: 58.5%, Yahoo: 54.8% and MSN: 17.8%).

The reason for not linking more to exact same page might be in the fact that they treat redirects as copies of the same document and only link to the one that they think is the original document.

Correlation

Correlation coefficients are calculated using Pearson's product-moment correlation coefficient.

Correlations for number of Wikipedia results (including 0 results)

	Google	Yahoo	MSN
Google	1	0.507 **	0.367 **
Yahoo	0.507 **	1	0.459 **
MSN	0.367 **	0.459 **	1

** Correlation is significant at the 0.01 level (2-tailed).

Correlations of highest ranks

	Google	Yahoo	MSN
Google	1	0.421 **	0.146 **
Yahoo	0.421 **	1	0.245 **
MSN	0.146 **	0.245 **	1

** Correlation is significant at the 0.01 level (2-tailed).

Correlations of exact Wikipedia links

	Google	Yahoo	MSN
Google	1	.689 **	.275 **
Yahoo	.689 **	1	.398 **
MSN	.275 **	.398 **	1

** Correlation is significant at the 0.01 level (2-tailed).

From these correlations we can observe that Google and Yahoo have high correlation ratio between themselves with MSN having much way of treating Wikipedia articles.

Discussion and further research

Above analysis shows that Wikipedia ranks highly and favorably in Google search and Yahoo search engine. On the other hand MSN search does not follow their example. It ranks Wikipedia more unfavorably or it does not include it all in top 10 results. Users using MSN search engine are thus less likely to use Wikipedia as source of their information.

Looking at correlations results we can also see that MSN has quite different ranking algorithm than Google and Yahoo. This probably means that it also ranks other pages differently. This can be seen as part of their strategy to be better search engine than their rivals and ranking algorithm can be one of their differencing factors. Obtained data in top 10 results could be used to more thoroughly explore this question.

Additional exploration of how search engines treat copies of same articles inside Wikipedia can be done. This analysis did not differentiate between a redirect and normal page but looking at results of comparison of exact URLs, we can see that search engines treat them differently.

Results can be further improved by improving methodology by adding additional source of queries as control variable. Extra differentiation of redirects pages could give better insights as would use of more complex sampling approaches. It should be possible to sample articles from within biggest categories to further see if there is a difference in ranks of covered topics. This could potentially be used to measure how informative are Wikipedia pages in regard to that specialized part of Internet community, as seen by search engines.

Conclusion

Observations of different blog authors were not without merit. There is indeed a high ratio of Wikipedia results for general search queries. The fact that they use Google as their search engine helped with their observations since Google ranks Wikipedia highly. Average user searching for topics that are well covered in Wikipedia can expect to often see Wikipedia results in top 10 results list. Especially when using Google or Yahoo search.

While analyzing data a significant difference of ranking algorithm was observed for MSN search in comparison to Google and Yahoo.

Sources

Carr, Nicholas G., Information central, http://www.roughtype.com/archives/2006/08/the_centralizat_1.php, August 2006, last accessed 11. October 2006

Carr, Nicholas G., Our new Delphic oracle, http://www.roughtype.com/archives/2006/08/the_oracle_of_w.php, August 2006, last accessed 11. October 2006

Lehtonen, Risto, Pahkinen, Erkki J., Practical Methods for Design and Analysis of Complex Surveys, 2004, John Wiley and Sons

Rubel, Steve, Study: Wikipedia Dominates Brand Search Results, http://www.micropersuasion.com//Brands%20and%20Wikipedia.pdf and http://www.micropersuasion.com/2006/09/study_wikipedia.html, September 2006, both last accessed 11. October 2006

Vehovar, Vasja, Kalton Graham: Vzorčenje v anketah (Monograph “Survey sampling” in Slovenian language), 2001, Faculty of Social Science.