

# Using User Reviews to Improve Search in Wikipedia

Yasser Ganjisaffar, Sara Javanmardi and Cristina Lopes  
School of Informatics & Computer Science  
University of California, Irvine  
Irvine, CA, USA  
{yganjisa, sjavanma, lopes}@ics.uci.edu

## ABSTRACT

The encyclopedic knowledge accumulated in Wikipedia is so large that one often uses search engines, to find information in it. In contrast to regular Web pages, Wikipedia articles are accompanied with history pages, categories and talk pages. The meta-data available in these pages can be analyzed to gain a better understanding of the content and quality of the articles. We analyze the quality of search results of the current major Web search engines (Google, Yahoo! and Live) in Wikipedia and discuss how the rich meta-data available in wiki pages can be used to provide better search results in Wikipedia. We investigate the effect of incorporating the extent of review of an article in the quality of rankings of the search results. The extent of review is measured by the number of distinct editors who have contributed to the articles and is extracted by processing Wikipedia’s history pages. Our experimental results show that re-ranking search results of the three major Web search engines using the review feature improves quality of their rankings for Wikipedia-specific searches.

## 1. INTRODUCTION

Using wiki technology, Wikipedia has become the largest online encyclopedia, used as a reference for encyclopedic knowledge [10]. Wikipedia, as a massive repository of knowledge, is most useful when its articles are well-organized and easily accessible. Web search engines have been successful in making the Web content accessible for a decade, and they succeed in searching Wikipedia too. However, there are differences that make wiki content distinguishable from traditional Web content. For example, Each Wikipedia article is accompanied with history, talk, and category pages. History pages comprise old revisions of the wiki text as well as the record of the timestamp and username of the contributor. Category pages provide semantic knowledge about the concept which is presented by the article. This meta-data can provide significant insight into the content of the article and can be used to provide higher quality search results to users.

The semantics of links that connect wiki pages can also be different from the semantics of the links in traditional Web pages. Whereas in Web documents an author can arbitrarily link his page to any other page, whether there is a topical relation or not, a significant fraction of links in Wikipedia point to semantically related content [9, 8]. In addition, some links are inserted automatically by Wikipedia’s registered bots<sup>1</sup> serving particular purposes. In aggregate, Wikipedia’s links can not simply be interpreted as votes for authoritativeness of the target page as it is in traditional Web pages. For example, articles representing years like “2008” have a large number of in-links compared to most of the other articles. This issue suggests that link-based ranking algorithms like PageRank [12] might be less effective in the domain of Wikipedia.

Based on these differences, it is necessary to study the current status of search in Wikipedia and possible improvements that can be achieved by considering the meta-data available in Wikipedia.

A recent study on Wikipedia [13] shows that high-quality articles in Wikipedia benefit from higher number of edits and distinct contributors. In addition, the editors of Wikipedia can be considered as a sample of the interests of the general population. Therefore, articles edited by a lot of reviewers are probably the most popular articles (i.e. most searched for) in the general population. We propose a review-based ranking algorithm to improve quality of search in the domain of Wikipedia. We show that the quality of the rankings by the current major Web search engines can be improved incorporating the proposed heuristic in their ranking schemes.

The contributions of this work are twofold. First, the empirical study of search performance by the three major search engines in Wikipedia is presented. Second, the review-based heuristic proposed here not only results in considerable improvements for the two least-performing search engines, but it suggests that the future of search on the Web will need to take into account the social activities that are now taking place in it.

The remainder of this paper is organized as follows. In Section 2, we analyze the differences in search performance in Wikipedia between the three major Web search engines, namely Google, Yahoo! and Live. Section 3 shows how those search engines are affected by adding the additional review-

<sup>1</sup>[http://en.wikipedia.org/wiki/Wikipedia:Bot\\_policy](http://en.wikipedia.org/wiki/Wikipedia:Bot_policy)

based heuristic. Section 4 presents related work, and Section 5 concludes the paper.

## 2. SEARCH IN WIKIPEDIA

Based on the differences between Wikipedia content and the general Web content, we set out to study the effectiveness of major Web search engines in searching Wikipedia. In the remainder of this section, we present the current state of the three major Web search engines, namely Google, Yahoo! and Live search, in terms of retrieval effectiveness in Wikipedia and freshness of their index.

To compare the effectiveness of the rankings of search results, we use the evaluation metric called Normalized Discounted Cumulative Gain (NDCG for short) [6]. This metric measures the usefulness, or *gain*, of a document based on its position in the results list. The gain is accumulated from the top of the results list to the bottom with the gain of each result discounted at lower ranks.

We asked seven graduate students in different majors to use our interface for searching Wikipedia, and to label search results as highly relevant (HR), relevant (R), and irrelevant (IR). Students were asked to search for both special topics related to their major and general topics. All of the three search engines have application programming interfaces (API) that allow programs to submit queries and get the search results. After a query is submitted through our interface, it is submitted to Google, Yahoo!, and Live APIs. Queries are appended with “site:en.wikipedia.org” to restrict domain of search to English Wikipedia.

For presenting search results to users for labeling, we used the *pooling* method [7]. Query is submitted to the three search engines and the top 10 results from each search engine are added to a pool. Duplicates in search results are removed and the final set of results is *randomly* presented to user for labeling. A total of 240 queries were submitted and 3,410 results were labeled.

Figure 1 shows NDCG values for positions 1 through 10 for Google, Yahoo!, and Live search engines. Higher NDCG values show higher quality rankings. NDCG values are shown for different number of positions. For example, the NDCG value at position 4 shows the normalized gain when considering only the top 4 results of the search engine. For the top 1 search results, the three search engines have approximately similar gains. However, Live outperforms the rest for the top 2 to top 10 results. Given that Live search treats Wikipedia pages different than other pages (e.g, see [1]), it seems that it is using some Wiki-specific information in ranking its search results.

Our study also shows the diversity of the algorithms used by the three major search engines to retrieve and rank relevant results to users’ queries, which makes their results different even on Wikipedia as a small fraction of the Web. Table 1 shows that the three search engines agree on the top search result for 61.7% of the queries. It also shows percentage of common pages in top- $k$  results returned by the the search engines. Considering the top 10 results of each search engine, only 23.4% of results are common among them.

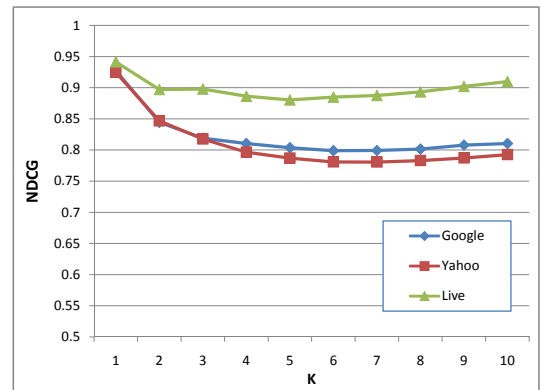


Figure 1: Quality of search results of major Web search engines in domain of Wikipedia

	Top 10	Top 5	Top 1
Google & Yahoo!	34.3%	35.7%	66.5%
Google & Live	37.9%	38.1%	69.6%
Yahoo! & Live	36.5%	37.3%	65.1%
All	23.4%	24.9%	61.7%

Table 1: Percentage of common pages in top- $k$  search results returned by the three major Web search engines.

## 3. IMPROVING SEARCH IN WIKIPEDIA

Unlike many traditional models of knowledge and publishing, which attempt to limit content creation to a relatively small group of approved editors in order to exercise strong quality control, Wikipedia articles are published without strict prior quality checking. The open editing model of Wikipedia allows users to *review* and edit previously contributed content by other users in order to improve its quality. Because of the radical openness in Wikipedia, the peer reviewing process is scaled up orders of magnitude compared to peer reviewing process of scholarly articles. The results of the study in [13] show that Wikipedia’s featured articles<sup>2</sup> benefit from higher number of edits and distinct contributors. In addition, the editors of Wikipedia can be considered as a sample of the interests of the general population. Therefore, articles edited by a lot of reviewers are probably the most popular articles (i.e. most searched for) in the general population.

In order to analyze the usefulness of the extent of review in ranking Wikipedia articles, we processed the English Wikipedia. We crawled the English Wikipedia domain (en.wikipedia.org) in January 2009. The crawlers were programmed to discard redirected articles (about 35% of the articles) and eliminate HTML templates from the downloaded articles. After elimination of duplicates, a total of 2,448,558 articles were indexed. We downloaded the dump of the Wikipedia history

<sup>2</sup>Featured articles are considered to be the best articles in Wikipedia. Before being listed as featured, each article is reviewed by some experts familiar with the subject for accuracy, neutrality, completeness, and style ([http://en.wikipedia.org/wiki/Featured\\_Article](http://en.wikipedia.org/wiki/Featured_Article)).

released in October, 2008<sup>3</sup> and extracted number of distinct editors contributed in each article.

We assigned an integer review score between 0 (lowest) and 10 (highest) to each of the articles using a logarithmic scale. The reason for using a logarithmic scale for review scores is that as the number of editors increases, adding more editors would have less effect on quality. For example, there is not much difference between an article having 400 editors and another having 401 editors. Also, we do not want an article with 400 editors to have a score which is 200 times of the score of an article with 2 editor. The reason for quantizing review scores is that integer review scores between 0 and 10 can be stored in one byte of memory. But for storing floating point numbers at least 4 bytes in needed.

The exact process for calculating review scores is as follows: we first find the maximum and minimum number of editors an article has in our data set. The article for “George W. Bush” has the maximum number of editors which is 13, 198. There are lots of articles with only 1 editor. Therefore our minimum value would be 1. Our next step would be to quantize the numbers in range [1, 13198] using a logarithmic scale. In our experiments, we tried different bases for the logarithmic scale and 3 resulted in the most uniform distribution. Therefore, we transform our range to the logarithmic scale with base 3 which make the range [0, 8.64]. Since we want to have 11 review scores (from 0 to 10), we divide this range to 11 equally sized ranges. For each of the articles, we first get the logarithm of the number of its editor (in base 3) and based on which of the 11 ranges contains this number, we assign it to the corresponding review score.

There is a positive correlation between distinct number of contributors (review scores) and content quality, and also between content quality and PageRank of articles [13, 11]. Hence, we expect to see a positive correlation between PageRank and review scores. To verify this hypothesis, during our crawl of Wikipedia, we stored its internal link structure. A total of 256, 304, 639 links were extracted. We used the PageRank algorithm to find the relative importance of articles according to their in-links. The PageRank values converged after 29 iterations. We scaled the PageRank values to a value between 0 and 10 using a logarithmic scale. Figure 2 shows the average review scores of articles as a function of PageRank. Review scores increase with the rise of PageRank values which supports the positive correlation between review scores and article quality.

Table 2 lists the top English Wikipedia articles based on PageRank and Review scores. The list of top articles based on PageRank values is dominated with country pages and Wikipedia’s special pages like “Wikipedia Commons” and “Geographic Coordinate System”. These pages have a large number of in-links compared to other pages. For example, in most of the articles about United States cities, the latitude and longitude of the city is specified in the page and there is a link to “Geographic Coordinate System” article near these values. Some of these links are inserted by Wikipedia’s registered bots that make automated edits in articles. However, all of the edits by bots are counted only once in the list of

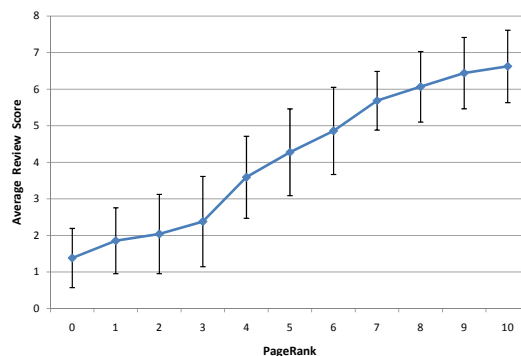


Figure 2: Average review scores of the articles as a function of PageRank values

Most Reviewed	United States, 2007, United Kingdom, Canada England, New York City, World War II, 2006 India, Germany, World War I, Wikipedia The Beatles, Adolf Hitler, George W. Bush Ronald Reagan, Jesus, Bill Clinton, Wiki Harry Potter, Led Zeppelin, Hurricane Katrina Albert Einstein, Michael Jackson, Star Wars The Simpsons, Xbox 360, Wii, Metalica
High PageRank	United States, 2007, United Kingdom, Canada England, France, Wikipedia Commons Geographic Coordinate System, New York City World War II, 2006, India, Germany, Russia Scotland, Norway, 2008, English Language, 2005 Australia, New Zealand, Europe, London, Spain Poland, China, Italy, Sweden, Netherlands, Japan Brazil, California, 2004, Census, Public domain

Table 2: Top most reviewed and high PageRank articles in English Wikipedia

most reviewed articles.

We conducted an experiment to analyze if adding extent of review as a new feature can help improve rankings of the three search engines for Wikipedia articles. We divided the 240 queries data set labeled during our previous experiments to a 120 queries training set and a 120 queries test set. We used the training set to train a support vector machine (SVM) classifier<sup>4</sup> in order to see if  $doc_i$  should be ranked higher than  $doc_j$ , for the query  $q$  according to the following features:

- Position of  $doc_i$  among the top 10 results returned by search engine.
- Position of  $doc_j$  among the top 10 results returned by search engine.
- Difference between positions of  $doc_i$  and  $doc_j$ .
- Review score of  $doc_i$ .

<sup>3</sup><http://download.wikimedia.org/enwiki/20081008/>

<sup>4</sup><http://svmlight.joachims.org/>

- Review score of  $doc_j$ .
- Difference between review scores of  $doc_i$  and  $doc_j$  on an exponential base<sup>5</sup>.

Figure 3 plots the average NDCG for positions 1 to 10 for the three search engines on our test set and compares results with those gained by using the SVM classifier to rank search results. The results show that incorporating the review score improves the quality of ranking; but it is more apparent for Google and Yahoo! search engines. In case of Live search engine, review scores improve the quality of ranking only when we consider more than 7 positions in search results. Since review score is an extremely simple feature to calculate, totalling the number of editors of the article, it is very promising to see such improvement.

#### 4. RELATED WORK

While it is difficult to measure Wikipedia’s overall quality in a definitive way, some studies have tried to assess it in several ways. Some characteristics such as factual accuracy [4] and credibility [3], have been used to compare small samples of Wikipedia articles to their parallel articles in other reputable sources.

In [11], the number of edits and unique editors to an article were suggested as metrics for quality. Built on this study, Wilkinson and Huberman [13] proposed a statistical analysis of Wikipedia and showed that featured articles are distinguishable from the rest by high number of edits and distinct number of contributors. They also verified the validity of this observation for articles with different age and visibility.

In [5], Hu *et al.* have proposed a framework that re-ranks Wikipedia search results considering article quality. They have developed two quality measurement models, Basic and PeerReview. Article quality is derived based on co-authoring data gathered from articles’ edit history. According to their experimental results, compared with Google, Wikiseek<sup>6</sup> and Wikipedia’s internal search engine, rankings generated based on their quality models are less accurate. However, they showed some improvements in rankings of Wikipedia’s search engine and Wikiseek by combining their PeerReview Model to their rankings. The PeerReview model requires processing the whole history of each article for assessing its quality. Consecutive revisions should be compared to find out who has added each word and who has reviewed it. Given the high dynamics of Wikipedia which has resulted in more than 615 million revisions as of this writing, this model is computationally expensive and does not seem to scale well.

#### 5. CONCLUSION & FUTURE WORK

Using Wiki technology, Wikipedia has become a massive online knowledge base. Its articles become most useful when they are well-organized and easily accessible. Although current search engines help users locate content in Wikipedia,

<sup>5</sup>Review scores are scaled on a logarithmic scale. Therefore, it is necessary to measure the difference of review scores on an exponential basis.

<sup>6</sup>A Wikipedia search engine which is no longer active as of 2008

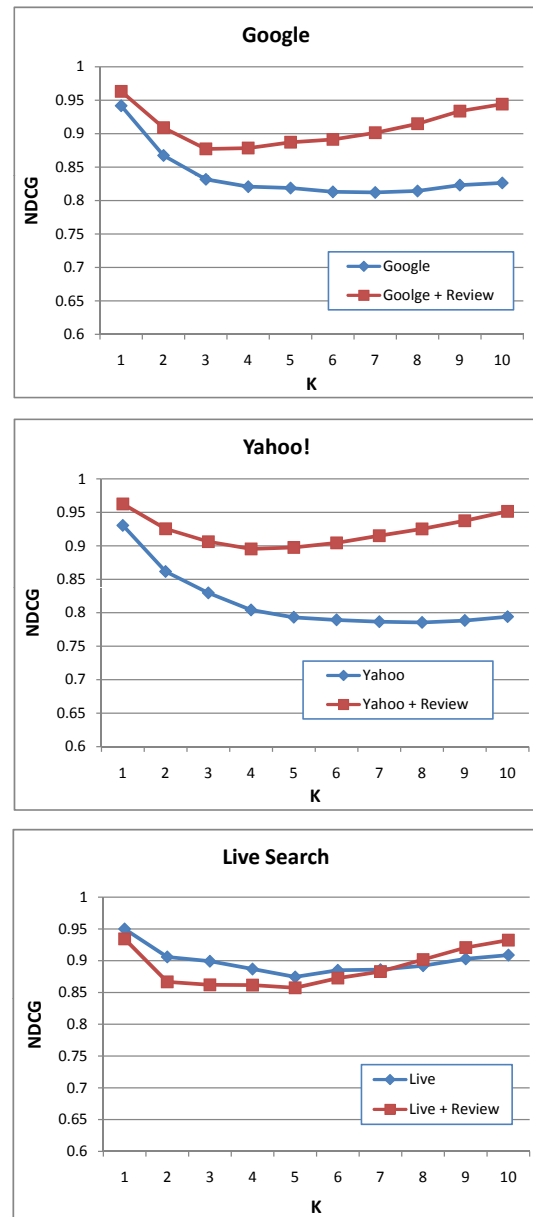


Figure 3: Quality of search results of major Web search engines in domain of Wikipedia and impact of re-ranking their results using the review-based ranking algorithm

there are some new features introduced by wiki technology that can be taken into account for ranking the results.

In this paper, we explored the differences between wiki pages and traditional web pages and studied the effectiveness of search in the domain of Wikipedia between three major Web search engines: Google, Yahoo! and Live Search. Based on the studies on crowdsourcing systems and hypothesis of wisdom of crowd, we introduced a very simple review-based ranking algorithm to rank Wikipedia articles in search results. We described results of our experiments that show the number of distinct editors of an article that can be easily extracted from history of articles can improve ranking of articles presented to users.

More sophisticated information can be extracted from history, category, and talk pages. This meta-data can be used to better evaluate quality of articles and provide better rankings of Wikipedia search results. In the future, we plan to develop a custom search engine for Wikipedia to use the additional information available in Wikipedia to provide better search results. For example, Category information can be used to provide categorized search results. Organizing search results allows users to focus on items in categories of interest rather than having to browse through all the results sequentially. Results of the study in [2] show that users are 50% faster at finding information when search results are organized into categories.

Overall, our study suggests that as the Web becomes more structured around social activities, the heuristics for search will likely have to adapt.

## 6. REFERENCES

- [1] The official blog of the live search team: Wikipedia gets big. Electronically.
- [2] H. Chen and S. Dumais. Bringing order to the web: automatically categorizing search results. In *CHI '00: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 145–152, New York, NY, USA, 2000. ACM.
- [3] T. Chesney. An empirical examination of wikipedia’s credibility. *Firstmonday*, 11(11), November 2006.
- [4] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, December 2005.
- [5] M. Hu, E.-P. Lim, A. Sun, H. W. Lauw, and B.-Q. Vuong. On improving wikipedia search using article quality. In *WIDM '07: Proceedings of the 9th annual ACM international workshop on Web information and data management*, pages 145–152, New York, NY, USA, 2007. ACM.
- [6] K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 41–48, New York, NY, USA, 2000. ACM.
- [7] K. S. Jones and C. van Rijsbergen. Report on the need for and provision of an “ideal” information retrieval test collection. Technical Report British Library Research and Development Report 5266, University of Cambridge, Computer Laboratory, 1975.
- [8] J. Kamps et al. Is wikipedia link structure different? In *WSDM '09: Proceedings of the second ACM international conference on Web search and data mining*, New York, NY, USA, February 2009. ACM.
- [9] J. Kamps and M. Koolen. *Advances in Information Retrieval*, chapter The Importance of Link Evidence in Wikipedia, pages 270–282. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2008.
- [10] B. Keim. News feature: Wikimedia. *Nature Medicine*, 13:231–233, March 2007.
- [11] A. Lih. Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. In *Proceedings of the International Symposium on Online Journalism*, pages 16–17, 2004.
- [12] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.
- [13] D. M. Wilkinson and B. A. Huberman. Assessing the value of cooperation in wikipedia. *Firstmonday*, 2007.