

# Review-based Ranking of Wikipedia Articles

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**—Wikipedia, the largest encyclopedia on the Web, is often seen as the most successful example of crowdsourcing. The encyclopedic knowledge it accumulated over the years is so large that one often uses search engines, to find information in it. In contrast to regular Web pages, Wikipedia is fairly structured, and articles are usually 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 discuss how the rich meta-data available in wiki pages can be used to provide better search results in Wikipedia. Built on the studies on “Wisdom of Crowd” and the effectiveness of the knowledge collected by a large number of people, 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 contributed to the articles and is extracted by processing Wikipedia’s history pages. We compare different ranking algorithms that explore combinations of text-relevancy, PageRank, and extent of review. The results show that the review-based ranking algorithm which combines the extent of review and text-relevancy outperforms the rest; it is more accurate and less computationally expensive compared to PageRank-based rankings.

**Keywords**-Wikipedia; Search; Ranking

## I. INTRODUCTION

Social networks and social network analysis in particular is a research paradigm which tries to unravel patterns of social relationships across various individuals in a social context [1], [2]. Wiki software facilitates a case where social relationships are established over a domain of social actions such as acceptance or rejection of a contribution. For example in the case of Wikipedia, the wiki facilitates a collaborative document editing effort relying on the contribution of a large number of users in a concurrent system. This enables an effective combination of the ground knowledge about a topic (Wisdom of Crowd) in the most recent revision of the article. In that sense, the current revision of an article in Wikipedia is the outcome of a community process involving certain social interactions embedded in the content modification, used as a mean of expressing them [3], [4].

Wikipedia’s model of knowledge creation, which allows anyone to enter and edit content, has enabled its rapid expansion: since its inception in 2001, Wikipedia has grown to encompass 11.9 million articles in 265 languages gener-

ated from 615 million edits by 15 million contributors<sup>1</sup>. Its current position as the 7th most visited website<sup>2</sup> confirms its usefulness and popularity.

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, the special features introduced by wiki technology make search in the domain of Wikipedia different from traditional Web content. Some of the main differences are summarized as follows:

- *Meta-data information.* 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.
- *Wikipedia link structure.* The semantics of links that connect wiki pages can 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 [5], [6]. In addition, some links are inserted automatically by Wikipedia’s registered bots<sup>3</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 might be less effective in the domain of Wikipedia.
- *Multi-ownership.* Wikipedia articles are created and edited by co-contributors. Thus, reliability of a wiki

<sup>1</sup>[http://meta.wikimedia.org/wiki/List\\_of\\_Wikipedias](http://meta.wikimedia.org/wiki/List_of_Wikipedias)

<sup>2</sup><http://www.alexa.com>

<sup>3</sup>[http://en.wikipedia.org/wiki/Wikipedia:Bot\\_policy](http://en.wikipedia.org/wiki/Wikipedia:Bot_policy)

page is not necessarily correspondent the reputation of a single entity. The open editing model of Wikipedia allows users to contribute to any article regardless of their attributions or expertise.

Based on these differences, it is important for Web search engines to adapt themselves to wiki environments. In this paper, we investigate how the additional meta-data available in Wikipedia can be used to improve the quality of ranking of the search results in domain of Wikipedia.

The remainder of this paper is organized as follows. Section II presents related work. In Section III, we shows how extent of review can be incorporated in ranking Wikipedia articles. Section IV explains a method for comparing different ranking algorithms. This method is used in Section V to compare the effectiveness of the review-based ranking algorithm and PageRank-based ranking for searches in domain of Wikipedia. Finally, Section VI concludes the paper.

## II. RELATED WORK

Web search has been studied as a classic information retrieval problem. Link analysis techniques such as PageRank [7] and HITS [8] measure the popularity of Web pages based on their interlinking structure, and this popularity is used in ranking search results to yield better search performance. The PageRank score of a page is derived from the scores of pages linking to that page. In HITS algorithm, each page is assigned a hub and an authority score. A page deserves a high hub score when it provides links to authoritative pages and high authority score for being referenced by good hub pages. High PageRank, hub and authority pages are likely to be high quality pages [9], [10].

Besides PageRank and HITS, numerous metrics have been studied in literature to measure Web page quality [11], [9], [12], [10]. In particular, in [10], Zhu and Gauch have studied six metrics for assessing Web page quality, namely *currency*, *availability*, *information-to-noise ratio*, *authority*, *popularity* and *cohesiveness*, and found that incorporating quality metrics generally improved search effectiveness. However, these proposed quality metrics may not work for Wikipedia because most Wikipedia articles follow similar page design and offer equal accessibility.

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 [13] and credibility [14], have been used to compare small samples of Wikipedia articles to their parallel articles in other reputable sources.

In [15], the number of edits and unique editors to an article were suggested as metrics for quality. Built on this study, Wilkinson and Huberman [16] proposed a statistical

analysis of Wikipedia and showed that featured<sup>4</sup> 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.

The closest work to our study is presented in [17] and [18]. 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>5</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.

## III. EXTENT OF REVIEW OF ARTICLES

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.

Despite Wikipedia's success, we know little about why it has been so effective. One possibility is that having many contributors results in higher quality and less biased articles [19]. The benefits of aggregating judgments from many people have been observed since at least 1907, when Galton showed that averaging independent judgments of many observers estimated the weight of an ox at a county fair better than experts could [20]. Other recent studies show that a broad spectrum of human activities requiring critical decisions illustrate the greater reliability of judgments by crowds than by experts [21], [22]. A recent study on Wikipedia [16] shows that high-quality articles in Wikipedia benefit from higher number of edits and distinct contributors. Other studies [23], [21] on prediction in crowdsourcing systems show that the average of predicted scores by the crowd becomes more reliable as the size of the crowd increases.

<sup>4</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)).

<sup>5</sup>A Wikipedia search engine which is no longer active as of 2008

Table I  
DISTRIBUTION OF PAGERANK IN ENGLISH WIKIPEDIA ARTICLES

| PageRank | Count     |
|----------|-----------|
| 0        | 1,063,113 |
| 1        | 510,052   |
| 2        | 536,364   |
| 3        | 258,853   |
| 4        | 59,321    |
| 5        | 15,495    |
| 6        | 3,737     |
| 7        | 1,368     |
| 8        | 211       |
| 9        | 36        |
| 10       | 8         |

Some interpret this fact by the “law of large numbers” in which the mean of a sample of independent observations from a given population approaches the population mean as the sample size increases [24].

According to these observations, we expect the quality of the Wikipedia entries to improve as they go through iterations of edits by different users. We investigate the effect of incorporating the extent of review of articles in the quality of rankings of the search results. The extent of review is measured by the number of distinct editors contributed to the articles and is extracted by processing Wikipedia’s history pages.

There is a positive correlation between distinct number of contributors (review scores) and content quality, and also between content quality and PageRank scores [16], [15]. Hence, we expect to see a positive correlation between PageRank and review scores. To verify this hypothesis, we processed 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 stored the internal link structure of Wikipedia while crawling. 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. Table I shows the distribution of the PageRank values of articles.

We downloaded the latest dump of the Wikipedia history released in October, 2008<sup>6</sup> and extracted number of distinct editors contributed in each article. Figure 1 shows the distribution of number of editors of articles in log-log format.

We assigned a review score between 0 and 10 to each of the articles using a scale with logarithmic gaps. The exact

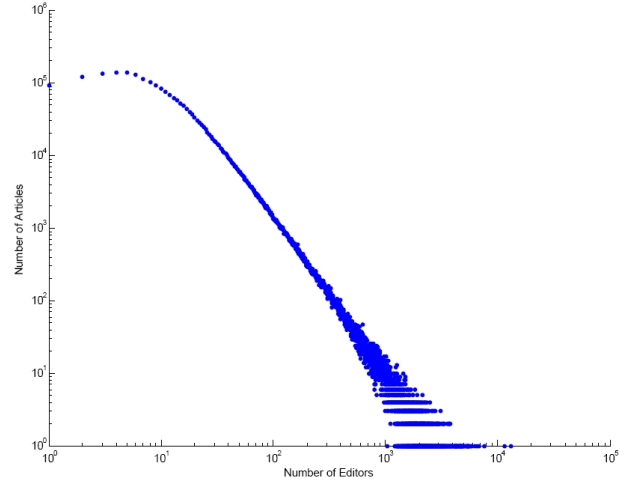


Figure 1. Distribution of Number of Editors of Wikipedia Articles

Table II  
DISTRIBUTION OF REVIEW SCORES IN ENGLISH WIKIPEDIA ARTICLES

| Number of Editors        | Review Score | Number of Articles |
|--------------------------|--------------|--------------------|
| $ed = 1$                 | 0            | 213,679            |
| $2 \leq ed \leq 3$       | 2            | 252,555            |
| $4 \leq ed \leq 7$       | 3            | 512,879            |
| $8 \leq ed \leq 19$      | 4            | 754,592            |
| $20 \leq ed \leq 55$     | 5            | 476,395            |
| $56 \leq ed \leq 163$    | 6            | 172,132            |
| $164 \leq ed \leq 489$   | 7            | 51,336             |
| $490 \leq ed \leq 1467$  | 8            | 12,881             |
| $1468 \leq ed \leq 4399$ | 9            | 2,066              |
| $4400 \leq ed$           | 10           | 43                 |

formula for calculating review scores is as follows:

$$s_{review} = \begin{cases} \frac{ed - ed_{min}}{ed_{max} - ed_{min}} < \frac{1}{b^{10}}, & 0 \\ \exists 1 \leq k \leq 9 \left( \frac{1}{b^{k+1}} \leq \frac{ed - ed_{min}}{ed_{max} - ed_{min}} < \frac{1}{b^k} \right), & 10 - k \\ \frac{1}{b} \leq \frac{ed - ed_{min}}{ed_{max} - ed_{min}}, & 10 \end{cases}$$

where,  $ed$  is the number of editors of the article;  $ed_{min}$  and  $ed_{max}$  are the minimum and maximum number of editors in Wikipedia articles;  $b$  is the scaling factor. The above formula assigns a review score to the article based on the relative position of  $ed$  between  $ed_{min}$  and  $ed_{max}$ . In our experiments, we tried different values for  $b$  and 3 resulted in the most uniform distribution. Table II shows the distribution of review scores of articles.

Figure 2 shows the average review scores of articles as a function of PageRank (bars show standard deviation). Review scores increase with the rise of PageRank values which supports the positive correlation between review scores and article quality.

Table III lists the top English Wikipedia articles based on PageRank and Review scores. The list of top articles based on PageRank values is dominated with articles about

<sup>6</sup><http://download.wikimedia.org/enwiki/20081008/>

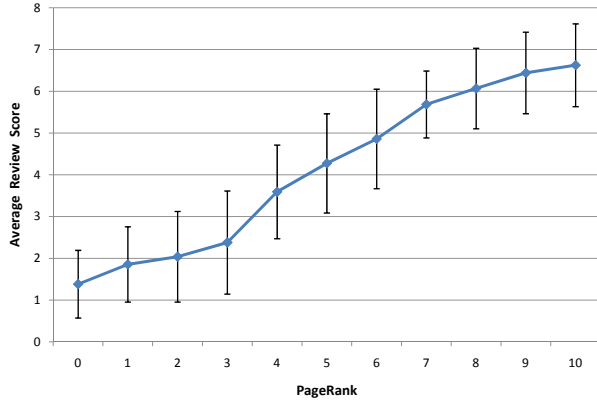


Figure 2. Average review scores of the articles as a function of PageRank values

Table III  
TOP MOST REVIEWED AND HIGH PAGERANK ARTICLES IN ENGLISH WIKIPEDIA

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

countries and some special articles like “Wikipedia Commons” and “Geographic Coordinate System” that have a large number of in-links compared to other articles. For example, in most of the articles about United States cities, the latitude and longitude of the city is specified in the article 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 most reviewed articles.

Modern search engines use tens of factors in ranking search results. Text relevancy, position of query terms in page, PageRank score, and website reputation are among those factors. Some of these factors might be meaningless for a Wikipedia-specific search engine. For example, since all of the Wikipedia articles are in a single domain, domain reputation can not be a factor in ranking articles. Other factors like PageRank might be computationally expensive and might not have the expected effectiveness (because of

the semantic differences between Web links and wiki links). Comparing the effectiveness of different ranking factors in wiki-specific searches can provide guidelines for developing wiki search engines. In the next section, we explain a method for comparing different ranking algorithms.

#### IV. COMPARING RANKING ALGORITHMS

Absolute judgments of users for the relevancy of search result can be used to compare the effectiveness of different rankings of Wikipedia articles. Although absolute judgments can be used in evaluation methods like NDCG [25] for detailed analysis of different methods, collecting such judgments for each search result of each query is expensive. Recent work has suggested the use of *preference judgments* [26], [27], [28]: given two documents, an assessor only expresses a preference for one over the other. Preference judgments are interesting because assessors can make preference judgments faster than absolute judgments on a graded scale [28]. Preference judgments can be explicit or implicit. In the explicit preference judgment, assessors are presented pairs of documents and requested to select one of them as more relevant. Implicit preference judgments are inferred from actions of users. For example, clickthrough data can be interpreted as implicit evidence of the relevancy of documents [29], [30].

A user is more likely to click on a link if it is relevant to query. This dependency between clicks and relevancy can be used in evaluating the ranking presented to users. However, there is also a dependency between clicks and presented ranking that makes evaluation harder: users are less likely to click on links low in the ranking, regardless of how relevant they are. Users typically do not scroll down the rankings too far to observe relevant links at the bottom.

Assume that after getting the results of query  $q$ , user has clicked on links 1, 3, and 6. While it is not possible to infer that the links 1, 3, and 6 are relevant on an absolute scale, it is much more plausible to infer that link 3 is more relevant than link 2 with probability higher than random. Assuming that the user scanned the ranking from top to bottom (as supported by eye-tracking studies [31]), he must have observed link 2 before clicking on link 3, making a decision to not click on it. Similarly, it is possible to infer that link 6 is more relevant than links 2, 4 and 5. This means that clickthrough data can be interpreted as the relative relevance judgments for the links the user browsed through. In [32], [31], [33], authors have conducted a series of experiments and concluded that this type of interpreting clickthrough data closely follow the human relevance judgments.

Based on the above interpretation of the clickthrough data, we extract a set of pairwise preferences from the clickthrough data and evaluate the accuracy of ranking algorithms based on the percentage of correctly ordered pairs. In the above example, preference pairs would be (3, 2), (6, 2), (6, 4) and (6, 5). A ranking algorithm that has ordered

documents as (1, 3, 2, 6, 4, 5) would have an accuracy of 75%; from the four preference pairs one of them, (6, 2), is presented as (2, 6) in the ordered list.

Ordering two items can be viewed as a binary classification problem in which the classifier decides to rank higher the first (+1) or the second item (-1). Therefore, a classifier can be trained on the pairwise preferences extracted from clickthrough data and used for ordering search results. Each pairwise preference ( $d_i, d_j$ ) is composed of two documents,  $d_i$  and  $d_j$ . We expect an accurate ranking algorithm to rank  $d_i$  above  $d_j$ . For each document  $d_i$  we have its PageRank and Review scores. Since each pairwise preference is extracted based on the clicks on results of a specific query, we can also extract the text-relevancy of the document to that query. To train the classifiers, we use the following set of features:

$$\begin{aligned} F_{(i,j)}^1 &= (s_i^{relevance} - s_j^{relevance}) \\ F_{(i,j)}^2 &= (s_i^{pagerank} - s_j^{pagerank}) \\ F_{(i,j)}^3 &= (s_i^{review} - s_j^{review}) \end{aligned}$$

Text relevancy ( $s^{relevance}$ ) is a value between 0 and 1 which is returned by the underlying indexing engine. We used Apache Lucene<sup>7</sup> as our index and search engine. Lucene assigns a text relevancy score to each of the documents in result set based on their textual relevancy to the query terms.

For each pairwise preference ( $d_i, d_j$ ), we generate the following training instances:

$$\begin{aligned} +1 : & \quad F_{(i,j)}^1 \quad F_{(i,j)}^2 \quad F_{(i,j)}^3 \\ -1 : & \quad -F_{(i,j)}^1 \quad -F_{(i,j)}^2 \quad -F_{(i,j)}^3 \end{aligned}$$

The first training instance states that  $d_i$  is more preferred than  $d_j$ ; the second one states that  $d_j$  is less preferred than  $d_i$ .

### Recording Clickthrough Data

Clickthrough data can be recorded with little overhead and without compromising the functionality of the search engine. It also does not add any overhead for the user compared to explicit user feedback. Whenever a user clicks on a result, an asynchronous message is sent to the server and user is forwarded to the clicked URL.

We developed an experimental search engine for Wikipedia named “Wikijoo”<sup>8</sup> to use as our basis for collecting clickthrough data. We configured Wikijoo to rank search results based on a linear combination of PageRank of articles and the text-relevancy:

$$S_i = \alpha \times s_i^{relevance} + (1 - \alpha) \times s_i^{pagerank}$$

<sup>7</sup>http://lucene.apache.org/

<sup>8</sup>Available online at: http://www.wikijoo.org

where,  $\alpha$  is a parameter to weigh text-relevancy in the combined score. We set  $\alpha$  to 0.5. Given that the interpretation of clickthrough data described in Section IV is not affected by the ranking of results presented to users [32], [33], the choice of  $\alpha$  does not affect our evaluations.

We asked a group of seven graduate students in different majors to use Wikijoo for one week and recorded their queries and clicks. Figure 3 shows a screenshot of how results are presented to users.

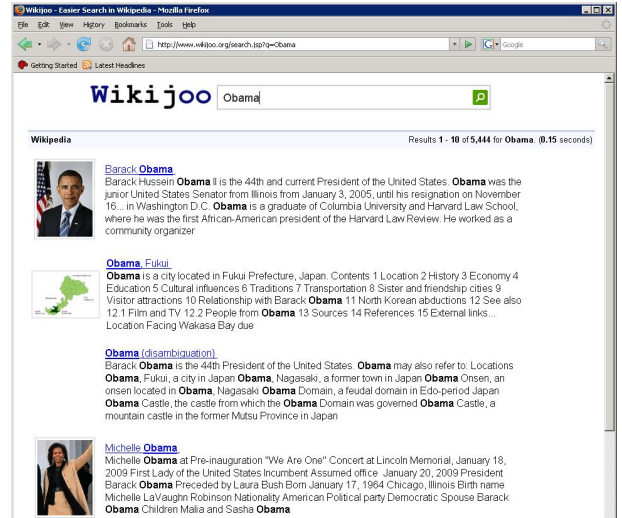


Figure 3. A screenshot of how results are presented in Wikijoo

## V. EXPERIMENTAL RESULTS

A total of 1,917 queries were recorded in Wikijoo. From this, only 70% resulted in clicks. Most of the queries which were not followed by clicks are either those containing spelling errors or those in which users decided to reformulate their queries after seeing results. Figure 4 shows the distribution of clicks on top 10 search results in Wikijoo, and compares them with distribution of clicks on Google’s search results as reported in [31]. More than 65% of the clicks in Wikijoo are on the first result. This value is significantly different from the 44% of clicks on first result of Google. Also average clicks per query in Wikijoo is 0.7, while this number is 0.9 for Google search results. We believe that these differences are because of the encyclopedic nature of Wikipedia’s content; there is typically one relevant article for each concept. For example, a Web Search engine might find thousands of relevant pages related to query “Barack Obama”, but there is a single page dedicated to this concept in Wikipedia while other pages are far less relevant to this concept.

We used SVM-light<sup>9</sup> to train binary SVM classifiers that use different set of features for ranking results. Table IV shows the results based on 10-fold cross validation.

<sup>9</sup>http://svmlight.joachims.org/

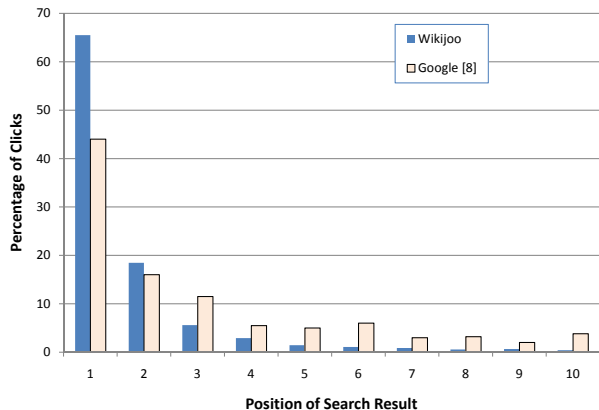


Figure 4. Distribution of clicks on Wikijoo and Google results

The accuracy of the first ranker, which uses only the difference between relevancy scores, is 85.27%. The next two rankers use only PageRank and Review scores for ranking results. They both have a better accuracy than the first ranker. Note that, before passing articles to the ranking algorithm they are filtered based on whether they have been relevant to the query or not.

The best accuracy is achieved in  $C_7$  which uses all of the features. With a small difference after that, is the  $C_6$  ranker. The interesting thing about this ranker is that by combining Review scores with the base text–relevancy based ranker ( $C_1$ ) we have gained 8.8% in accuracy which is a significant improvement. On the other hand, adding PageRank scores to the base text–relevancy based ranker improves its accuracy by 7.5%. This shows that review scores are a better choice in ranking Wikipedia articles, compared to PageRank scores.

Given that Review scores are calculated based on the number of distinct editors of the articles and this historical data can be easily extracted from Wikipedia database, this improvement is gained with very little overhead. Compared to this, calculation of PageRank of articles is computationally expensive. All of the links between pages must be extracted and stored at crawl time. The PageRank algorithm itself is also an iterative algorithm which is computationally expensive on large graphs (more than 256 million links in our data set). In our experiments, calculation of PageRank score of pages finished after 29 iterations and this required 124 minutes of computation on our machine<sup>10</sup>. Adding this amount of time to the overhead of extracting and storing links at crawl time shows the complexity of PageRank algorithm. It is interesting that the review–based ranking method which ranks articles based on a fundamental characteristic of Wiki environments (extent of review of articles) and comes at almost no computation overhead outperforms the PageRank–based ranking algorithm.

<sup>10</sup>2 Quad–core processors and 8GB RAM

| Name  | Features Used                      | Accuracy |
|-------|------------------------------------|----------|
| $C_1$ | Text–Relevancy                     | 85.27%   |
| $C_2$ | PageRank                           | 91.17%   |
| $C_3$ | Review                             | 91.94%   |
| $C_4$ | PageRank + Review                  | 92.00%   |
| $C_5$ | Text–Relevancy + PageRank          | 92.77%   |
| $C_6$ | Text–Relevancy + Review            | 94.07%   |
| $C_7$ | Text–Relevancy + PageRank + Review | 94.11%   |

Table IV  
EXPERIMENTAL ACCURACIES OF DIFFERENT CLASSIFIERS

## VI. CONCLUSIONS AND FUTURE WORK

An effective search engine for Wikipedia should locate relevant and high quality pages for given queries. We measured the extent of review of an article as an estimate of its quality and proposed a review–based ranking algorithm.

To evaluate the effectiveness of the review–based ranking algorithm, we compared it with other ranking algorithms that use combinations of text–relevancy, PageRank, and extent of review. The results show that the review–based ranking outperforms the rest with regard to accuracy and computational cost.

In the future work, we aim to analyze if adding quality measures like extent of review can improve the quality of rankings presented by external search engines like Google and Yahoo when searching in domain of Wikipedia.

### ACKNOWLEDGEMENT

This work has been partially supported by NSF grant OCI-074806.

### REFERENCES

- [1] S. Wasserman, K. Faust, and D. Iacobucci, *Social network analysis: Methods and applications (structural analysis in the social sciences)*. Cambridge University Press, 1994.
- [2] J. Scott, *Social network analysis: A handbook (2nd Ed.)*. London; Thousands Oaks, Calif.: SAGE Publications, 2000.
- [3] N. T. Korfiatis, M. Poulos, and G. Bokus, “Evaluating authoritative sources in collaborative editing environments,” *Online Information Review*, vol. 30, no. 3, pp. 252–262, 2006.
- [4] A. Kittur, E. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz, “Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie,” *World Wide Web*, vol. 1, no. 2, 2007.
- [5] J. Kamps and M. Koolen, *Advances in Information Retrieval*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2008, ch. The Importance of Link Evidence in Wikipedia, pp. 270–282.
- [6] 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: ACM, February 2009.

- [7] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford Digital Library Technologies Project, Tech. Rep., 1998.
- [8] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.
- [9] B. Amento, L. Terveen, and W. Hill, "Does 'authority' mean quality? predicting expert quality ratings of web documents," in *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2000, pp. 296–303.
- [10] X. Zhu and S. Gauch, "Incorporating quality metrics in centralized/distributed information retrieval on the world wide web," in *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2000, pp. 288–295.
- [11] T. Mandl, "Implementation and evaluation of a quality-based search engine," in *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*. New York, NY, USA: ACM, 2006, pp. 73–84.
- [12] Y. Zhou and W. B. Croft, "Document quality models for web ad hoc retrieval," in *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*. New York, NY, USA: ACM, 2005, pp. 331–332.
- [13] J. Giles, "Internet encyclopaedias go head to head," *Nature*, vol. 438, no. 7070, pp. 900–901, December 2005.
- [14] T. Chesney, "An empirical examination of wikipedia's credibility," *Firstmonday*, vol. 11, no. 11, November 2006.
- [15] 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*, 2004, pp. 16–17.
- [16] D. M. Wilkinson and B. A. Huberman, "Assessing the value of cooperation in wikipedia," *Firstmonday*, 2007. [Online]. Available: <http://arxiv.org/abs/cs.DL/0702140>
- [17] 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*. New York, NY, USA: ACM, 2007, pp. 145–152.
- [18] M. Hu *et al.*, "Measuring article quality in wikipedia: models and evaluation," in *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. New York, NY, USA: ACM, 2007, pp. 243–252.
- [19] A. Kittur and R. E. Kraut, "Harnessing the wisdom of crowds in wikipedia: quality through coordination," in *CSCW '08: Proceedings of the ACM 2008 conference on Computer supported cooperative work*. New York, NY, USA: ACM, 2008, pp. 37–46.
- [20] F. Galton, "Vox populi," *Nature*, vol. 75, pp. 450–451, 1907.
- [21] J. Surowiecki, *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday, 2004.
- [22] P. E. Tetlock, *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, 2005.
- [23] D. Reeves and D. Pennock, "How and when to listen to the crowd," Electronically. [Online]. Available: [http://www.overcomingbias.com/2007/02/how\\_and\\_when\\_to.html](http://www.overcomingbias.com/2007/02/how_and_when_to.html)
- [24] D. V. Schroeder, "Book reviews," *American Journal of Physics*, vol. 74, no. 3, March 2006.
- [25] 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*. New York, NY, USA: ACM, 2000, pp. 41–48.
- [26] B. Carterette and P. N. Bennett, "Evaluation measures for preference judgments," in *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2008, pp. 685–686.
- [27] B. Carterette, P. N. Bennett, and O. Chapelle, "A test collection for preference judgments," in *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2008.
- [28] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais, "Here or there: Preference judgments for relevance," in *Proceedings of ECIR 2008*, March–April 2008.
- [29] T. Joachims, "Optimizing search engines using clickthrough data," in *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2002, pp. 133–142.
- [30] W. W. Cohen, R. E. Schapire, and Y. Singer, "Learning to order things," *Journal of Artificial Intelligence Research*, vol. 10, pp. 243–270, 1999.
- [31] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay, "Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search," *ACM Transactions on Information Systems*, vol. 25, no. 2, p. 7, 2007.
- [32] F. Radlinski, M. Kurup, and T. Joachims, "How does click-through data reflect retrieval quality?" in *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge mining*. New York, NY, USA: ACM, 2008, pp. 43–52.
- [33] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," in *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2005, pp. 154–161.