On Improving Wikipedia Search using Article Quality

Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady W. Lauw and Ba-Quy Vuong
Centre for Advanced Information Systems
School of Computer Engineering
Nanyang Technological University
Singapore 639798
{hu0003un,aseplim,axsun,hady0002,vuon0001}@ntu.edu.sg

ABSTRACT

Wikipedia is presently the largest free-and-open online encyclopedia collaboratively edited and maintained by volunteers. While Wikipedia offers full-text search to its users, the accuracy of its relevance-based search can be compromised by poor quality articles edited by non-experts and inexperienced contributors. In this paper, we propose a framework that re-ranks Wikipedia search results considering article quality. We develop two quality measurement models, namely BASIC and PEERREVIEW, to derive article quality based on co-authoring data gathered from articles' edit history. Compared with Wikipedia's full-text search engine, Google and Wikiseek, our experimental results showed that (i) quality-only ranking produced by PEERREVIEW gives comparable performance to that of Wikipedia and Wikiseek; (ii) PEERREVIEW combined with relevance ranking outperforms Wikipedia's full-text search significantly, delivering search accuracy comparable to Google.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: [Information filtering]; H.3.5 [**Online Information Services**]: [Web-based services]

General Terms

Algorithms, Experimentation

Keywords

Wikipedia, collaborative authoring, quality-aware search

1. INTRODUCTION

1.1 Motivation

The Web has evolved from an information repository to a platform for information sharing and collaboration. With the increasing popularity of Web 2.0 applications (e.g., wikis, blogs, and social tagging), enormous amount of Web information are now contributed by individual Internet users and there is little control over the quality of such information. Given that user contributed information can have quality ranging from good to poor, it is therefore important for Web search engines to return *relevant* and *good quality* results as much as possible.

In this paper, we study quality-aware search for Wikipedia[1] articles. Wikipedia becomes our research focus for the following reasons:

- Wikipedia is the most successful wiki, in which more than 1.8 million articles (counting English articles alone, as of July 2007) have been contributed by thousands of contributors. Wikipedia has become the primary online knowledge sharing platform [20] and is currently among the top 10 most popular websites according to `Alexa.com`.
- Wikipedia articles are published without stringent prior quality checking and their contributors include non-experts and inexperienced users. Hence, not all articles are of desired equal quality.

Searching Wikipedia can be performed using either external search engines like Google or the Wikipedia built-in search engine. External search engines, however, suffer from several drawbacks[2]. For instance, external search engine may not index the latest version of Wikipedia articles on a real-time basis. Moreover, some search engines do not distinguish namespaces[3] used in Wikipedia, hence they do not give preference to encyclopedia proper pages over *Talk:* or *User:* pages, which are used for communication among contributors. On the other hand, the Wikipedia's full-text search engine also has its limitations as it ranks articles based on relevance only without considering their quality, which often results in less than expected performance[4].

1.2 Research Objectives and Contributions

In this research, we aim to design and evaluate quality-aware search methods for Wikipedia such that both *relevance* and *quality* are incorporated in ranking search results.

WIDM'07, November 9, 2007, Lisboa, Portugal.

[1] `http://en.wikipedia.org/`
[2] `http://en.wikipedia.org/wiki/Wikipedia:Searching`
[3] Namespace is used to compartmentalize Wikipedia pages. A full list of Wikipedia namespace can be found on page `http://en.wikipedia.org/wiki/Wikipedia:Namespace`
[4] `http://en.wikipedia.org/wiki/Wikipedia_talk:Searching#Wikipedia.27s_search_is_aweful.21`

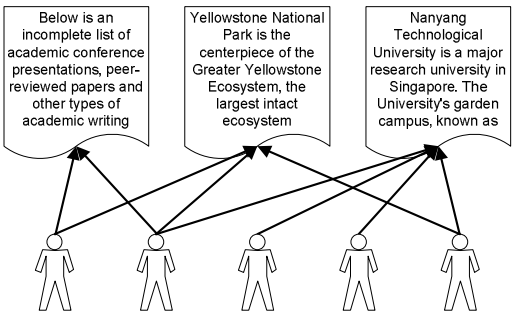


Figure 1: Articles, Contributors and their Association in Collaborative Authoring

The problem is challenging because determining article quality itself is not an easy task, even for human assessors. The difficulties can be attributed to:

- *Number of articles is huge.* Due to the exponential growth rate of Wikipedia, it is clearly a very laborious task to rate the quality of every article manually.
- *Articles are evolving.* Wikipedia is not static. As article content changes, so does their quality. This complicates the quality assessment task and may cause much more human efforts if the assessment is not done automatically.
- *Quality measurement is not trivial.* Quality itself is a subjective concept. As part of its effort to identify good quality articles, Wikipedia maintains a set of featured articles which are *well written*, *comprehensive*, *factually accurate*, *neutral* and *stable*, to name a few[5]. All these criteria are not easy to measure without careful study of an article content.

Our research therefore aims to design quality-aware search methods that determine article quality in Wikipedia automatically without interpreting the article content. We design our quality measurement models based on the collaborative nature of Wikipedia contributors. As shown in Figure 1, each article in Wikipedia may be edited by a set of contributors and each contributor may edit multiple articles. Our idea of calibrating article quality is based on determining the authority of their contributors and the mutual dependency between the article quality and contributor authority as stated below:

- **Quality**: An article has high quality if it is contributed by high authority authors.
- **Authority**: A contributor has high authority if he or she contributes high quality articles.

Our major contributions in this research can be summarized as follows:

- We propose a general framework to integrate the quality measurement models into the existing Wikipedia's full-text search.
- Based on this framework, we explore alternative search result re-ranking methods using both relevance and quality.
- We evaluate the ability of our proposed searching strategy in improving Wikipedia search results.

Although our work specifically addresses quality-aware search for Wikipedia articles, the same approach can be equally applied to other wikis and collaboratively authored Web content.

1.3 Paper Organization

The rest of this paper is organized as follows: Section 2 summarizes related work in Web search ranking and assessing Web document quality. We propose our quality-aware search framework in Section 3. Our quality measurement models are introduced in Section 4. We describe our experimental design in Section 5. Section 6 presents our experiment results and discussions. Finally, Section 7 concludes this paper.

2. RELATED WORK

Web search has been studied as a classic information retrieval problem, where coverage is used to evaluate a search engine [8, 13]. Link analysis techniques such as PageRank [17] and HITS [12] measure the popularity of Web pages based on their interlinking structure, and this popularity can be exploited in ranking search results to yield better search performance. The PageRank score $R(p)$ of a page p is derived from the scores of pages linking to p. With HITS, each page is assigned a *hub* score 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 page rank, hub and authority pages are likely to be high quality pages [2, 23].

Quality measure is subjective, and there is not yet a universal standard. Besides PageRank and HITS, numerous metrics have been studied in literature to measure Web page quality, including simple metric such as *document size* and many complicated metrics [2, 11, 16, 18, 22, 23]. In particular, Zhu and Gauch studied six metrics in assessing Web page quality, namely *currency*, *availability*, *information-to-noise ratio*[6], *authority*, *popularity* and *cohesiveness*, and found that incorporating quality metrics generally improved search effectiveness [16, 23]. In these studies, quality was treated independently from relevance, and better search results were achieved by considering both relevance and quality. Nevertheless, these proposed quality metrics may not work for Wikipedia because most Wikipedia articles follow similar page design and offer equal accessibility. Furthermore, many proposed metrics are subjective and require data supplied from sources outside Wikipedia [11, 18]. Studying the effectiveness of these metrics on Wikipedia is part of our future work.

Wikipedia's rich data content has also attracted growing interest in the research community [1, 3, 21]. Several research work closely related to ours include: evaluating article quality using metadata in article edit history [14], assessing article trustworthiness [21] and deriving user reputation in the context of evolving article revisions [1, 3]. Zeng et al [21] discussed a method to compute the trustworthiness of Wikipedia articles from a dynamic Bayesian network. They

[5] `http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria`

[6] *Information-to-noise ratio* is the ratio between the number of terms in a document after indexing and the raw size of the document (including HTML tags).

hypothesized that "the trustworthiness of the revised content of an article depends on the trustworthiness of: the previous revision, the authors of the previous revision, and the amount of text involved in the previous revision". They defined the trustworthiness of authors using Beta distributions according to 4 groups of users. Adler and Alfaro [1] discussed a method of computing reputation for Wikipedia users based on contribution survival in article edit history. Users have their reputation increased by longer preserved edits and reduced by their soon undone edits. Anthony et al [3] used an analogy of collective goods to describe Wikipedia articles. They analyzed the correlation among user registration status, participation level, and their contribution. [1] and [3] focused on profiling contributors rather than evaluating articles having uneven quality. Thus, the question of how to assess the quality of the large number of articles was left unanswered.

A search engine designed for Wikipedia known as Wikiseek[7] was recently launched, which is reported to enjoy high quality results and less spam [4]. Wikiseek utilizes user tagging and categorization information within Wikipedia to improve its search accuracy [4, 6]. It is not clear how quality has been considered in searching. As part of our study, the search performance of Wikipedia's internal search engine, Wikiseek as well as Google, will be compared with our proposed quality-aware search methods in Section 6.

Our pioneer work in measuring Wikipedia article quality was presented in [15]. In that work, we formalized the dependency between article quality and contributor authority as our *mutual reinforcement principle*; based on which, we proposed novel models, namely *basic* and *peer review*, to measure article quality in collaborative authoring. In this paper, we further improve *peer review* model to compute authority more accurately, which takes in review efforts made by each contributor. Also, we extend and apply article quality to search result ranking so as to examine the effectiveness of our proposed search methods.

3. PROPOSED SEARCH FRAMEWORK

The task of quality-aware search for Wikipedia is to locate *relevant articles* of *high quality* for a given query. Although quality is the main focus of this research, relevance remains the primary requirement for a search task. As shown in Figure 2, we propose a quality-aware search framework for Wikipedia that includes designated modules which derive quality scores for articles returned in searching for a query, and re-ranks these articles incorporating quality.

Given a query, the *Relevance Scoring Module* performs relevance search so as to return a set of candidate articles. Wikipedia's full-text search engine or other search engines customized for Wikipedia may be used for this module. From a set of search results, we construct an article *base set* in which the quality of every candidate article would be computed. There are more than one way to construct the *base set*. A simple strategy is to get a certain amount of top ranked articles from one or more search engines [12, 19].

Given an article base set, the *Quality Scoring Module* computes the quality for each article. In this paper, BASIC and PEERREVIEW are the two models designed for this purpose.

The *Re-Ranking Module* incorporates the relevance and quality of every article in the base set so as to give final

[7] http://www.wikiseek.com/

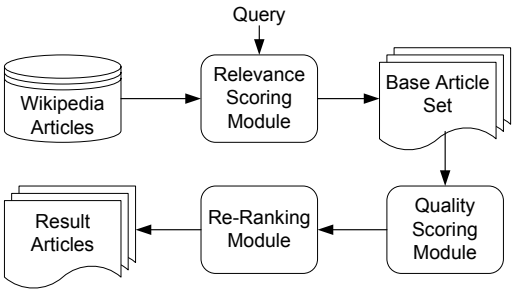


Figure 2: Quality-aware Search Framework

ranking of the search results. There could also be multiple ways to perform re-ranking. In general, a linear scheme to combine relevance and quality can be formulated as in Equation 1:

$$\bar{s}_i = \gamma \times s_i^{rel} + (1-\gamma) \times s_i^{qual} \quad (1)$$

where, $\bar{s}_i$ denotes the combined score of an article used in final ranking; s_i^{rel} and s_i^{qual} denote the relevance and quality scores of the article respectively; γ is a parameter to weigh relevance in the combined score, which takes values in the range $[0, 1]$.

This linear scheme offers flexibility in two aspects. The first is to determine s_i^{rel} and s_i^{qual}. These scores could take the computed relevance/quality values, or the assigned relevance/quality ranks, or some transformed scores of either. The second flexibility is to determine γ. Weight γ could be set at a fixed value for all search queries, or it could be determined in the search procedure adaptively. In this paper, we leave these options open and explore some of the alternatives in Section 5.3.

4. QUALITY MEASUREMENT MODELS

The novelty of our search framework lies in the use of a quality measurement model to derive article quality. As discussed in Section 2, article quality could be measured in numerous ways. A naive way to judge quality is by looking at the length of the article. We refer to such measurement model as the NAÏVE model [15]. As the name implies, this model could be easily fooled by very long articles artificially created as a means of vandalism[8]. For such cases, NAÏVE model will not be adequate and we will need a better way to differentiate good contributions from the poor ones.

We propose two models, namely BASIC and PEERREVIEW, to measure article quality based on the association between articles (word instances in articles) and their contributors derived from the edit history. The symbols and their semantics used in the formulation of our models are defined in Table 1[9]. It is worth pointing out that each word refers to a unique word instance in an article. Hence, identical words appearing at different positions within an article are different word instances in our definition as they may be authored by different contributors. For the same reason, articles do not share common words. The role a contributor plays in an article can be an *author*, a *reviewer* or both. Each word has its corresponding author and reviewer(s) which can be identified from article edit history (see Section 5.2).

[8] http://en.wikipedia.org/wiki/Wikipedia:vandalism

[9] In this paper, $|S|$ denotes the number of elements in set S.

Table 1: Symbol Semantics

Symbol	Semantic
A_j	denotes authority of a contributor u_j.
Q_i	denotes quality of an article a_i.
q_{ik}	denotes quality of a word w_{ik} in article a_i.
$w_{ik} \overset{A}{\leftarrow} u_j$	denotes u_j authored word w_{ik} in a_i. Each word has exactly one author.
$w_{ik} \overset{R}{\leftarrow} u_j$	denotes u_j reviewed word w_{ik} in a_i. A word may have zero or more reviewers. Reviewership is distinguished from authorship.
c_{ij}	denotes number of words u_j authored in a_i, i.e., $c_{ij} = \lvert\{w_{ik} \mid w_{ik} \overset{A}{\leftarrow} u_j\}\rvert$.

4.1 Basic Model

The BASIC model is designed to overcome the pitfalls of the NAÏVE model by adopting the assumptions introduced in Section 1.2 as the mutual dependency between article quality and contributor authority. More specifically, the BASIC model assumes that "good articles are contributed by high authority authors; and high authority authors write good articles". Formally, the BASIC model defines quality of article (Q_i) and authority of contributor (A_j) in Equations 2 and 3 respectively.

$$Q_i = \sum_j c_{ij} \times A_j \quad (2)$$

$$A_j = \sum_i c_{ij} \times Q_i \quad (3)$$

Equations 2 and 3 resemble the hub and authority definitions respectively. Instead of using links between articles, BASIC uses amount of contributions, i.e., c_{ij}'s, to determine the amount of authority values (quality values) that can be propagated to the quality values (authority values) of the authored articles (contributing authors).

4.2 PeerReview Model

Contributors do not only author content but also review articles. The way Wikipedia is designed encourages peer review on works among the contributors. Any peer reviewer can correct errors, rephrase sentences, or expand a whole paragraph. Corrected content improves the content's accuracy, while unchanged parts signify consensus among authors and reviewers. With this review mechanism, we can therefore assume that content reviewed and approved by high authority contributors should carry high quality. An article with much high quality content is therefore assumed to have high quality. This effectively gives another interpretation of the mutual dependency between article quality and contributor authority. That is, article quality is an aggregation of contribution from both high authority authors and high authority reviewers. This idea is realized by our PEERREVIEW model, as shown in Equations 4 and 5.

$$q_{ik} = \sum_{w_{ik} \overset{A}{\leftarrow} u_j \vee w_{ik} \overset{R}{\leftarrow} u_j} A_j \quad (4)$$

$$A_j = \sum_{w_{ik} \overset{A}{\leftarrow} u_j \vee w_{ik} \overset{R}{\leftarrow} u_j} q_{ik} \quad (5)$$

Equation 4 derives the quality of a word instance w_{jk} (denoted by q_{ik}) by summing the authorities of its author and reviewers. Equation 5 defines the authority of a contributor by summing the quality of words he/she has authored and reviewed. The intuition of counting on reviewed words in addition to authored words in calibrating contributor authority is based on the observation that good contributors are also those who consistently review vast amount of articles in Wikipedia. Quality of an article is then the aggregate quality of all its words, formally, $Q_i = \sum_k q_{ik}$.

The BASIC and PEERREVIEW models for a given set of articles and their contributors consist of a set of linear equations whose ranked solution can be obtained by iterative computation. Such iterative computation works by first assigning some initial values to the variables (i.e., quality and authority) in the equations, and iteratively updating the variables by applying the equations. This formulation and the convergence of the solution has been studied intensively in [7, 12]. In our experiments, we assume convergence when the delta changes of all variables between successive iterations are less than 10^{-6} [15, 19]. Our experiments have shown that both the BASIC and PEERREVIEW models converged in no more than 41 iterations.

5. EXPERIMENTAL DESIGN

The objective of our experiments is to examine the effectiveness of our proposed models in retrieving articles that are both relevant to the query and well composed so as to yield greater user satisfaction. To the best of our knowledge, there has not been any previous work on evaluating search performance on Wikipedia[10]. With no existing benchmark data set, we have therefore chosen to adopt a 20-query set and conducted user study on the query results for performance evaluation. In this section, we describe the query set used in our experiments, the search methods to be evaluated and the evaluation metrics.

5.1 Query Set and Article Base Set

We identified 20 queries that have been studied in Web search ranking [5, 12, 19], as shown in Table 2. This set consists of 10 single-term queries and 10 two-term queries. These queries carry fairly general meanings and are expected to have a few relevant Wikipedia articles each.

For each query, we obtained an article base set by feeding the query into three search engines, namely Wikipedia's internal search engine, Google and Wikiseek. All searches were done on 20 June, 2007. Among the results from these search engines, we considered only articles that fall into the Wikipedia default namespace, i.e., namespace = 0.

To avoid bias towards any single search engine, we took 500[11] top ranked articles from each search engine and combined them into the article base set for each query. Due to $\leq$ 500 search results and overlapping of results among search engines, the base set of a query ranged from 298 to 1,338 articles. The edit history as of 20 June, 2007 for each article in the base set was then acquired from Wikipedia.

5.2 Article Edit History Processing

Each revision of an article is known to be submitted by exactly one contributor. The author and reviewer(s) for each

[10]It came to our attention after the submission for review of this paper that INEX corpus included Wikipedia articles for XML retrieval in 2006.

[11]Maximum result limit set by Wikiseek.

Table 2: Query Set Statistics

q	Query	Base Set	Top 10	n_q^{HR}	n_q^{R}
1	abortion	997	44	12	25
2	alcoholism	298	46	8	17
3	basketball	1,190	78	14	55
4	bicycling	1,018	65	17	30
5	blues	1,072	76	5	14
6	cheese	750	63	8	18
7	genetics	790	66	11	33
8	java	1,002	58	18	9
9	movies	1,338	74	16	33
10	shakespeare	1,044	57	12	23
11	automobile industries	1,130	83	6	32
12	classical guitar	1,061	35	5	24
13	mutual funds	901	62	4	21
14	national parks	1,094	59	9	28
15	randomized algorithms	543	73	3	19
16	recycling cans	552	63	0	19
17	rock climbing	1,043	57	8	25
18	search engines	992	61	12	21
19	table tennis	948	62	3	21
20	vintage cars	890	103	4	19

Top 10: size of the union of top 10 results by all search methods.
$\mathbf{n_q^{HR}}$: number of HR-labeled articles.
$\mathbf{n_q^{R}}$: number of R-labeled articles.

word instance in the latest revision of each article were extracted from edit history by procedures described as follows:

- We extracted the lexicon for each revision with punctuation, stop words and Wikipedia's markup syntax removed. The relative order in word instances were retained for revision comparison.
- We performed *Diff* comparisons between the latest revision and every older revision in reverse-chronological order.
  - When a word instance in the latest revision is found to have existed in an older revision, the contributor of the older revision is added as a reviewer of the word instance;
  - When a word instance is found to be missing in all older revisions, the last added reviewer of that word instance is assigned as the author.

Table 3 summarizes the association statistics in our base set articles and their contributors for the 20 queries after edit history processing.

5.3 Methods to be Evaluated

We evaluate three types of search methods, namely (a) relevance-only, (b) quality-only and (c) average-rank methods. Our proposed BASIC and PEERREVIEW quality models are used in (b) and (c). We also include NAÏVE, which simply ranks articles by length, as one of the quality measurement models for comparison. The abbreviations of these methods are given in Table 4.

Relevance-only search methods: The search methods using relevance-only approach do not consider article quality measured by our models in the final result ranking for each query. In our experiments, we denote these methods as WIKI, Google and Wikiseek respectively, and use them as the baselines for comparison.

In the case of WIKI, because of Wikipedia's redirect mechanism, multiple alternative titles might lead to one single article[12]. Therefore, we resolve such search result by taking the true title that was associated with the encyclopedia content and omitting content duplicates. Removing such redirecting article from search results gives some unfair advantage to WIKI since the original full-text search engine tends to rank the true titles poorer than their redirect aliases.

Quality-only search methods: These are search methods that only use quality values by NAÏVE, BASIC and PEERREVIEW models to rank all articles in the base sets. We include these search methods to find out if quality can dominate relevance in search result ranking.

Average-rank search methods: These methods derive a combined score for each result from the search engine by the linear combination scheme defined in Equation 1, using relevance rank, normalized quality rank and $\gamma = 0.5$[13]. For example, if an article was ranked 1st by WIKI and 6th by BASIC among all articles in the base set, it would first receive a normalized quality rank of 3[14], and then get a combined score of $0.5\times1+(1-0.5)\times3 = 2$ in Wk+B. All search results from each search engine are then re-ranked using combined scores.

5.4 User Assessment for Search Results

As part of the evaluation on search methods, we conducted user assessment on the top 10 articles returned by each search method. That is, for each query, we take the union of top 10 search results returned by every search method (see Table 2). User assessment is then conducted on this union of articles. 10 article results per query is considered reasonable because both Google and Wikiseek show 10 results-per-page by default; and, most Web users are not likely to access search results beyond the first page [9].

User assessment on the search results of the 20 queries was based on two judgements, namely *relevance* and *quality*. For each query, we adopted three labels as the overall judge for each top 10 result, namely Highly Recommended (HR), Recommended (R), and Not Recommended (NR). The decision rules adopted in assigning labels are summarized as the following:

Relevant	Quality	Label
yes	high	**HR**
yes	moderate	**R**
yes	poor	**NR**
no	–	**NR**

As shown in Table 2, n_q^{HR} and n_q^{R} represent the numbers of labeled articles within the top 10 results returned by any

[12]The redirected-to article contains true encyclopedia proper content. In contrast, the redirecting articles provide the redirect mechanism to encyclopedia proper articles.

[13]We have attempted obtaining γ from the search module in an adaptive manner, i.e., deriving γ from the relevance scores of the top 10 search results using minimum, average, or exponential weighted sum. However, the resulting performance did not show much difference from that using $\gamma = 0.5$. Those approaches are therefore not reported in this paper.

[14]$3 = |\{result\ set\}_{Wk}| \times \frac{6}{|\{base\ set\}|}$, assuming that $|\{result\ set\}_{Wk}| = 500$ and $|\{base\ set\}| = 1000$. Normalization helps to avoid bias introduced by quality rank since the size of base set is generally twice that of each search engine result set.

Table 3: Average Article Base Set Statistics

Entity Association		min	max	avg	std dev
# authors	per article	1	363.4	21.4	34.9
# articles	per author	1	146.5	1.4	2.8
# words	per article	4.2	9,251.9	551.1	838.6
	per author	1	13,419.4	36.7	216.5
	per contribution	1	6,843	26	116.1
	per reviewer	1	67,319.8	803.7	1,874.9
# reviewers	per article	1	592.5	31.4	56.8
# articles	per reviewer	1	196.1	2	4.6

Table 4: Method Abbreviation

Method Type	Abbreviation
relevance-only	WIKI (Wk)
	Google (Gl)
	Wikiseek (Ws)
quality-only	NAÏVE (N)
	BASIC (B)
	PEERREVIEW (P)
average-rank	Wk+{N,B,P}
	Gl+{N,B,P}
	Ws+{N,B,P}

search methods (i.e., types (a), (b) and (c)). These labeled results are hence regarded as the ground truth in our evaluation.

5.5 Evaluation Metrics

We adopt an evaluation metric called **Normalized Discounted Cumulative Gain at top k** ("NDCG@k" for short) to compare the effectiveness of search methods. NDCG was first defined by Jarvelin et al [10], to measure search accuracy considering the multiple degrees of assessment in the search results. In our query result ranking problem, HR-labeled articles are considered more relevant than R-labeled articles, and we thus adopt NDCG to distinguish these two different degrees of relevance.

$$G_q(k) = \frac{1}{N_q} \sum_{p=1}^{k} \frac{2^{r(p)} - 1}{\log(1+p)} \quad (6)$$

As shown in Equation 6, NDCG@k for a query q is computed by summing up the gains from position $p = 1$ to $p = k$. Given the rank position p, $r(p)$ is an integer representing the amount of reward given to the article at p. In our case, $r(p) = 2$ when the pth ranked article is labeled HR. Similarly, $r(p) = 1$ for R-labeled article, and $r(p) = 0$ for NR-labeled article.

The term N_q is a normalization factor for query q, derived from a perfect ordering of top k articles that would yield a $G_q(k)$ of 1. Intuitively, the perfect ordering ranks all HR-labeled articles before all R-labeled articles. Formally,

$$N_q = \sum_{p=1}^{k} \frac{2^{s(p)} - 1}{\log(1+p)} \quad (7)$$

where,

$$s(p) = \begin{cases} 2 & \text{if } 1 \le p \le n_q^{HR} \\ 1 & \text{if } (n_q^{HR}+1) \le p \le (n_q^{HR} + n_q^{R}) \\ 0 & \text{otherwise} \end{cases}$$

6. EXPERIMENT RESULTS

In the first set of experiments, we compare the performance of all search methods using NDCG@k metric averaged over 20 queries. Figure 3 plots the average NDCG@k for k from 1 to 10.

6.1 Relevance/Quality-only Search Methods

Figure 3(a) compares the three relevance-only search methods (WIKI, Google and Wikiseek) and the three quality-only search methods (NAÏVE, BASIC and PEERREVIEW). This figure shows: (i) Google's top 10 performance is the best among all of six search methods; (ii) the quality-only PEERREVIEW always perform better than BASIC and NAÏVE for all k from 1 to 10; and (iii) none of the three quality-only methods is able to outperform the relevance-only methods especially at small k, despite that PEERREVIEW's performance is comparable to that of WIKI and Wikiseek at larger k. The last observation suggests that, there were articles of high quality but not relevant to the queries that were included in our base set, and they were ranked to the top by the quality-only methods. Hence, this result is within our expectation.

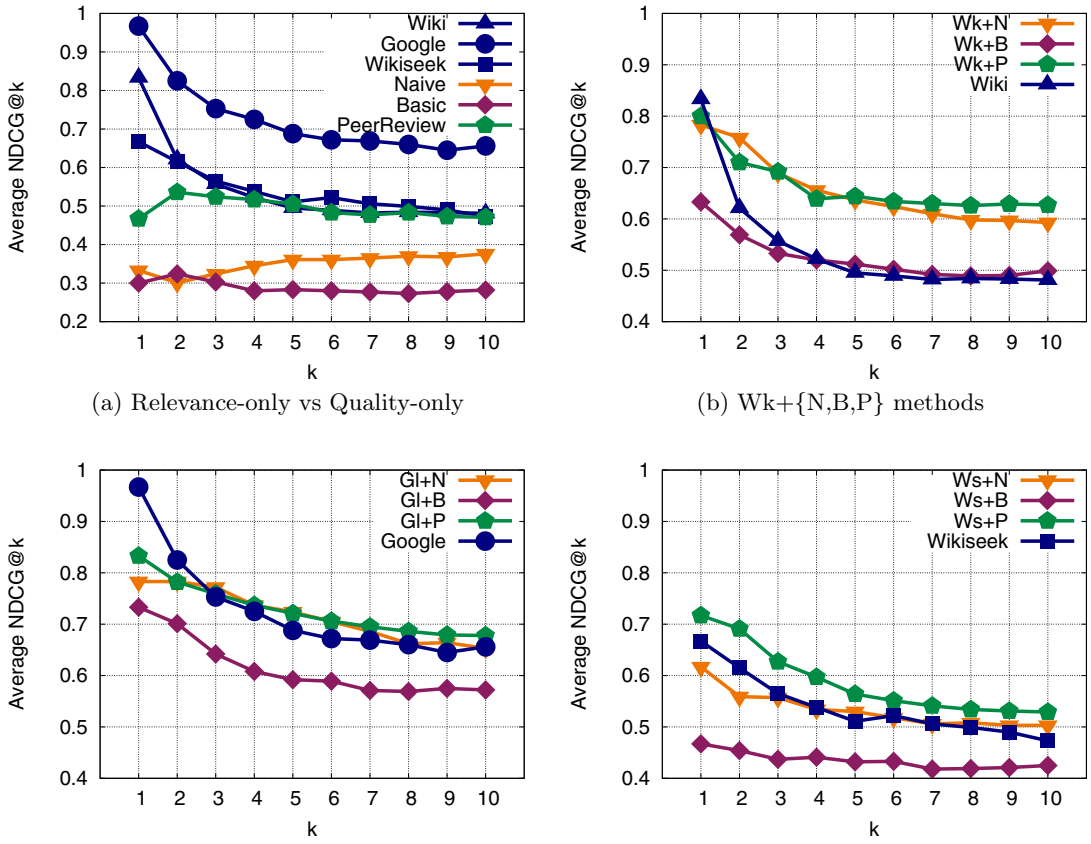
6.2 Average-rank Search Methods

As shown in Figures 3(b), 3(c) and 3(d), by considering both relevance and quality, we expect the average-rank search methods to give better performance than their respective relevance-only methods. As shown in Figure 3(b), Wk+P and Wk+N delivered significantly better search results than WIKI for almost all k values. At $k = 10$, which equals the default number of results on the first search result page, the NDCG values of Wk+P and Wk+N outperform that of WIKI by 31.2% and 24.2% respectively. Our paired t-test using p-value of 0.05 indicates that such improvement is significant from the 20 queries, i.e., having p-values at 2.2×10^{-4} and 5.4×10^{-3} respectively. The good NDCG@k performance of Wk+P and Wk+N suggests that, by incorporating article quality returned by the PEERREVIEW and NAÏVE models, we are able to give final search result ranking better than the original WIKI. This performance is comparable with Google's top 10 results[15].

As shown in Figures 3(c) and 3(d), Gl+P and Ws+P methods give higher NDCG@k than Google (for $k \ge 4$) and Wikiseek (for k from 1 to 10) respectively. However, not all average-rank search methods seem to enjoy the same performance gain over relevance-only methods using Google and Wikiseek. Gl+B and Ws+B generally do not perform better than their respective relevance-only search methods. This result suggests that the new articles in the top 10 search results and the new ordering of search results introduced by the quality models do not always produce better top results ranking. There is clearly more room for improvement in these sets of results as shall be discussed in Section 6.3.

There are two other interesting observations. The first observation is that Wk+N and Gl+N performed only slightly worse than Wk+P and Gl+P respectively. This suggests that, for those articles involved in our queries, article length could be an effective quality measure. This observation may

[15]Our paired t-test using p-value of 0.05 suggested non-significant difference in NDCG@$k = 10$ from Wk+P's 0.618 to Google's 0.646 for the 20 queries.



(a) Relevance-only vs Quality-only

(b) Wk+{N,B,P} methods

(c) Gl+{N,B,P} methods

(d) Ws+{N,B,P} methods

Figure 3: Average NDCG@k at Varying k

be useful since NAÏVE model is simple to implement. However, in the *open and free* Wikipedia, the NAÏVE model can be easily abused by malicious edits.

The second observation lies in the continual poor performance by search methods using BASIC model. In Figure 3(b), Wk+B is only able to perform better than WIKI by a small margin when $k \geq 5$. The degrade in performance of using article quality measured by BASIC is seen in Figures 3(c) and 3(d). We found that BASIC's performance was largely affected by the article-contributor association structure of the base set. When the distribution of c_{ij} in the article base set were skewed (i.e., there is one very large c_{ij} in the base set, and all other contributions contain much fewer words), the largest c_{ij} would draw bias towards the corresponding contributor u_j and the article a_i. In our experimented queries, we found that the biased u_j often caused not only the corresponding a_i but also a number of other articles authored by u_j being ranked to the top. Query 'automobile industries' was one such example: the largest c_{ij} was contributed by user *Vogensen* to article *Itapurange* containing 789 words, which was significantly larger than other contributions with 28.4 words on average. BASIC ranked this article the 1st by quality and this user the 1st by authority. Since *Vogensen* had authored 122 other articles about municipalities in Goiás state, Brazil, these articles also received good quality ranks in BASIC. Due to the irrelevance of these articles to the query and their considerable amount (i.e., 10.9% of base set), the performance of search methods using BASIC were largely worsened.

This observation pointed to us a future work direction as to develop an intelligent strategy to construct base set. Interestingly, PEERREVIEW model generally performs well for all 20 queries, regardless of the different authorship and reviewership distributions among base sets. This observation gives us more confidence in the robustness of PEERREVIEW model.

In summary, this set of experimental results shows that the quality-only search methods do not outperform relevance search methods, and average-rank methods using PEERREVIEW are most promising although their margin of improvement over Google is so far very little.

6.3 Further Discussion

The improvement in performance of average-rank methods on Google search results was seen very little. We suspect the cause to be some positive correlation between features used by Google in result ranking and the quality we have computed in our models. In an earlier study, Lih argued that press citations to certain Wikipedia articles had increased Web traffic to them and the quality of these articles had been improved consequently [14]. In other words, more densely linked articles are more likely to attract larger contributor population, and potentially have more experts involved, thus enjoy better chances to develop high quality.

To first investigate the effect of backlink on search-enigne performances, we extracted the backlinks[16] for the top 10 articles for each query returned by Google as well as the other two search engines. Our investigation suggests that among the three search engines, Google gives the largest average backlink counts for its results at every top 10 ranks. And, the rank 1 result from Google has much larger average backlink counts than the subsequent ranked results, i.e., 739.9 links on average for rank 1 results, compared with 140.6 to 242.2 links on average for other top 10 results.

In an attempt to map the correlation between backlink counts and article quality measured by our PeerReview model, we computed the Pearson correlation coefficient between these two quantities for top 10 search results from Google. We found the coefficients for 20 queries range from -0.316 to 0.978, having average of 0.498 and median of 0.608. On the whole, the positive correlation between backlink counts and article quality is supported by majority of our queries. This reveals that the little improvement in performance using average-rank methods on Google results is therefore not a surprise.

7. CONCLUSIONS

As Wikipedia continues to grow, the needs for automatic quality assessment techniques and quality-aware search are seen more evident. In this research, we have proposed a search framework that produces ranked results incorporating both relevance and quality. Basic and PeerReview are the two quality measurement models we have developed to measure article quality based on co-authorship among Wikipedia contributors. In the experiments we conducted, our quality-aware search methods incorporating PeerReview quality model have shown encouraging performance on the Wikipedia's full-text search results. However, there was room for improvement when re-ranking Wikipedia articles returned from other search engines.

Having identified some reasons for the models' behavior, we believe the following directions would lead us to interesting explorations and more promising results: (i) comparing the effectiveness of our proposed models with the measurement of other quality metrics [21, 23]; (ii) developing an intelligent strategy for constructing the article base set for each query rather than directly from search engine results; (iii) modeling contributors' expertise more accurately such that expertise-related authority could help retrieve more relevant results.

8. ACKNOWLEDGMENTS

This work was supported in part by A*STAR Public Sector R&D, Project Number 062 101 0031. We thank our anonymous reviewers for their comments and suggestions.

9. REFERENCES

[1] B. T. Adler and L. de Alfaro. A content-driven reputation system for the Wikipedia. In *Proc. of WWW'07*, pages 261–270, May 2007.

[2] B. Amento, L. Terveen, and W. Hill. Does "authority" mean quality? predicting expert quality ratings of Web documents. In *Proc. of SIGIR'00*, pages 296–303, July 2000.

[3] D. Anthony, S. Smith, and T. Williamson. Explaining quality in internet collective goods: Zealots and good samaritans in the case of Wikipedia, November 2005. Retreived online. `http://web.mit.edu/iandeseminar/Papers/Fall2005/anthony.pdf`.

[4] M. Arrington. Wikipedia search engine WikiSeek launches, January 2007. Published online. `http://www.techcrunch.com/2007/01/16/wikipedia-search-engine-wikiseek-launches/`.

[5] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the Web. In *Proc. of WWW'05*, pages 613–622, May 2005.

[6] S. Gilbertson. SearchMe launches Wikiseek, a Wikipedia search engine, January 2007. Published online. `http://blog.wired.com/monkeybites/2007/01/wikiseek_launch.html`.

[7] G. H. Golub and C. F. V. Loan. *Matrix Computations*. The Johns Hopkins University Press, 2715 North Charles Street, Baltimore, Maryland 21218-4363, 1996. 3rd edition.

[8] V. N. Gudivada, V. V. Raghavan, W. I. Grosky, and R. Kasanagottu. Information retrieval on the World Wide Web. *IEEE Internet Computing*, 1(5):58–68, September–October 1997.

[9] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: a study of user queries on the Web. *ACM SIGIR Forum*, 32(1):5–17, April 1998.

[10] K. Jarvelin and J. Kekalainen. IR evaluation methods for retrieving highly relevant documents. In *Proc. of SIGIR'00*, pages 41–48, July 2000.

[11] J. Jeon, W. B. Croft, J. H. Lee, and S. Park. A framework to predict the quality of answers with non-textual features. In *Proc. of SIGIR'06*, pages 228–235, August 2006.

[12] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, September 1999.

[13] S. Lawrence and C. L. Giles. Searching the World Wide Web. *Science*, 280(5360):98–100, April 1998.

[14] A. Lih. Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. In *Proc. of the 5th International Symposium on Online Journalism*, April 2004.

[15] E.-P. Lim, B.-Q. Vuong, H. W. Lauw, and A. Sun. Measuring qualities of articles contributed by online communities. In *Proc. of WI'06*, pages 81–87, 2006.

[16] T. Mandl. Implementation and evaluation of a quality-based search engine. In *Proc. of HYPERTEXT'06*, pages 73–84, August 2006.

[17] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web, November 1999. Technical Report, Stanford University Database Group. `http://dbpubs.stanford.edu:8090/pub/1999-66`.

[18] D. M. Strong, Y. W. Lee, and R. Y. Wang. Data quality in context. *Communications of the ACM*, 40(5):103–110, May 1997.

[19] P. Tsaparas. Using non-linear dynamical systems for Web searching and ranking. In *Proc. of PODS'04*, pages 59–70, June 2004.

[20] M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller, and R. Studer. Semantic Wikipedia. In *Proc. of WWW'06*, pages 585–594, May 2006.

[21] H. Zeng, M. A. Alhossaini, L. Ding, R. Fikes, and D. L. McGuinness. Computing trust from revision history. In *Proc. of the 2006 International Conference on Privacy, Security and Trust*, October–November 2006.

[22] Y. Zhou and W. B. Croft. Document quality models for Web Ad Hoc retrieval. In *Proc. of CIKM'05*, pages 331–332, October–November 2005.

[23] X. Zhu and S. Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web. In *Proc. of SIGIR'00*, pages 288–295, 2000.

[16] `http://developer.yahoo.com/search/siteexplorer/V1/inlinkData.html`