Measuring Article Quality in Wikipedia: Models and Evaluation

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 has grown to be the world largest and busiest free encyclopedia, in which articles are collaboratively written and maintained by volunteers online. Despite its success as a means of knowledge sharing and collaboration, the public has never stopped criticizing the quality of Wikipedia articles edited by non-experts and inexperienced contributors. In this paper, we investigate the problem of assessing the quality of articles in collaborative authoring of Wikipedia. We propose three article quality measurement models that make use of the interaction data between articles and their contributors derived from the article edit history. Our Basic model is designed based on the mutual dependency between article quality and their author authority. The PeerReview model introduces the review behavior into measuring article quality. Finally, our ProbReview models extend PeerReview with partial reviewership of contributors as they edit various portions of the articles. We conduct experiments on a set of well-labeled Wikipedia articles to evaluate the effectiveness of our quality measurement models in resembling human judgement.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: [Web-based services]; H.4 [**Information Systems Applications**]: [Miscellaneous]

General Terms

Algorithms, Experimentation

Keywords

Wikipedia, article quality, collaborative authoring, authority, peer review

CIKM'07, November 6–8, 2007, Lisboa, Portugal.
Copyright 2007 ACM 978-1-59593-803-9/07/0011 ...$5.00.

1. INTRODUCTION

Aiming at facilitating collaboration and information sharing, Wikipedia[1], *The Free Encyclopedia*, has evolved into the most popular wiki site worldwide[2]. Since its first launch in January 2001, Wikipedia has grown to approximately 7 million articles in 251 languages, among which more than 1.7 million are from the English Wikipedia[3]. The number of registered Wikipedians has grown to more than 4 million by early this year. The English Wikipedia alone has been doubling the number of its articles every year from 2002 to 2006[4].

The success of Wikipedia has to be attributed to the quality of its content [5, 8, 20, 23, 27]. A recent sampling, conducted by ***Nature*** revealed that the scientific entries in Wikipedia are of quality comparable to those in the more established Encyclopædia Britannica [8].

However, like many other open and free websites, Wikipedia has its fair share of quality problems. Other than vandalism and spamming, Wikipedia articles are not of uniformly good quality [21, 28]. One example is a hoax inserted into the Wikipedia article for John Seigenthaler, Sr., a well known journalist, linking him to the Kennedy assassinations. Although the hoax was finally detected and removed, it raised great concerns among the Wikipedia users and underlined the importance of quality assurance in Wikipedia.

1.1 Motivation

In this research, we aim to design models to measure quality of Wikipedia articles, which is an essential step in quality assurance. The ability to calibrate article quality brings about numerous benefits.

- Firstly, it helps readers to identify articles that are of good quality.
- Secondly, knowing which articles are of poor quality enables Wikipedia contributors to focus on the articles that require improvements and those needs to be replaced.
- Finally, article quality can be factored into searching and browsing strategies to improve the existing Wiki-

[1] `http://www.wikipedia.org/`
[2] Traffic rank on `http://www.alexa.com/`, as of April 2007.
[3] `http://en.wikipedia.org/`
[4] `http://en.wikipedia.org/wiki/Wikipedia`

pedia search engines and browsers, as well as other applications consuming Wikipedia information.

Determining the quality of articles in Wikipedia is not an easy task to human users, though there have been some serious attempts [6, 28]. The difficulties can be attributed to:

- *Large number of articles.* It is clearly a very laborious process for contributors to assess the quality of all Wikipedia articles, not mentioning that the number of articles keeps growing at an exponential rate [4].
- *Wide range of subject topics.* As the articles cover different topics, it requires experts from different disciplines to judge the quality. Unfortunately, such experts are not always available because they are usually busy outside Wikipedia due to their expertise.
- *Evolving content in the articles.* Wikipedia is never static. Every edit operation performed on an article may well affect its quality. This evolving nature therefore adds further complexity to the quality issues.
- *Varying contributor background.* Wikipedia contributors come from different geographic regions and diverse cultural backgrounds. Asking diverse reviewers to manually judge quality of articles may bring in non-uniform and subjective notions of quality.
- *Abuses.* Being a heavily accessed website, Wikipedia becomes a prominent target of different types of abuses, e.g. hoax, vandalism, spam, to name a few. Guarding articles from abuses is therefore a challenge.

Our objective is to develop quantitative measurement models to determine the quality of articles in Wikipedia with minimal human interpretation on the article content. We believe that by not having to check article content, more efficient quality checking can be conducted on Wikipedia, saving much effort of contributors in content creation. This work, however, is not about replacing the existing quality checking mechanisms in Wikipedia. Quality checking and quality measurement actually complement each other. Without the former, contributors will not be able to collaboratively produce good quality articles. Furthermore, quality measurement may also require interaction data produced by quality checking to calibrate article quality.

1.2 Summary of Contributions

In our approach, we observe the dependency between the quality of contribution to article content and the authority of contributors. We design quality measurement models that make use of the interaction data among articles and their contributors to produce quality ranking.

We summarize our research contributions as follows:

- We develop novel quality measurement models based on the interaction data gathered about the articles and their contributors in collaborative authoring. These are known as our BASIC, PEERREVIEW and PROBREVIEW models;
- We conduct a series of experiments to evaluate and compare the performance of models using a real article set from Wikipedia. We evaluate the rankings produced by our models against human manual judgement using NDCG metric commonly adopted in IR. The experimental results shows promising accuracy;
- In the experiments, we also analyze the effect article length in assessing quality. We explore the option of combining both authority and length features in deriving article quality ranking.

While our research is mainly designed for Wikipedia, it can be adopted and extended to measure quality of articles in other wikis and other websites that support collaborative editing[5]. This is possible because our proposed quality measurement models are built using the interaction data maintained by most wiki's. With the advent of Web 2.0, wikis and similar websites are expected to increase in number, hence making the quality measurement problem more critical, and quality measurement models like ours more relevant.

1.3 Paper Organization

The subsequent discussion of this paper is organized as follows. We summarize previous academic studies on evaluating Wikipedia in Section 2. We introduce our quality measurement models in Section 3, accompanied by our discussion on computational issues. Section 4 presents our experimental design and results. Finally, Section 5 concludes this paper.

2. RELATED WORK

Wikipedia has recently attracted growing interest in the research community [2, 5, 6, 16, 29]. Several research works closely related to ours were devoted to evaluating metadata to benchmark article status [16], assessing content trustworthiness [29] and user reputation [2, 5]. There were also other works on Wikipedia that focused on inter-article link analysis [1, 23], semantic relatedness [18, 24] and collection evolution [4, 26].

Lih's discussion [16] is among the very first on evaluating Wikipedia article in a systematic manner. He proposed a method to judge article quality based solely on metadata from the article edit history. Statistic features such as *rigor* (total number of edits) and *diversity* (total number of unique authors) were argued to be indicators of "level of good standing". He experimentally estimated the median values of these two features, which were then used to benchmark high quality articles. He also showed that citations from other established media has driven public attention directly to certain articles of Wikipedia, improving their quality subsequently.

Zeng et al. [29] modeled the trustworthiness of Wikipedia articles in a dynamic Bayesian network ('DBN' for short). Based on revision history, they 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". To define their DBN, they approximated the trustworthiness of Wikipedia authors as Beta distributions corresponding to 4 general user groups, namely administrators, registered users, anonymous users, and banned users. Their experiments using articles under Geography category of the English Wikipedia showed marginal lead by the mean

[5] A list of wiki sites can be found at WikiIndex, `http://wikiindex.org/`.

trustworthiness of featured articles[6] over the mean trustworthiness of clean-up articles[7].

Anthony et al [5] described Wikipedia articles as a form of Internet collective goods. They contributed a set of hypothesis on the correlation among user registration status, participation level, and the quality of their contribution. Adler and Alfaro assessed reputation gains for Wikipedia users based on content survival in revision history. They proposed a *chronological* method, in which each user gains their stimulated amount of reputation upon every arrival of new revisions. The cumulative reputation of an author is determined by how long his/her edited content could survive in terms of time span (*text survival*) and number of revisions (*edit survival*). The longer-lived edits would gain more reputation for their authors; those edits that only sustain in a short while in history would gain negative reputation for their contributors. Their experiments on the French and Italian Wikipedia has shown that: "changes performed by low-reputation authors have a significantly larger-than-average probability of having poor quality and being undone". They also reported low recall on 'new comers' in their *chronological* approach. Nevertheless, investigations in [5] and [2] focused on characterizing contributors rather than a horizontal comparison among the articles. Thus, the question of how to assess the uneven quality of the massive amount of articles was left unanswered.

We adopt a *fixed-point* data-driven approach in measuring article quality. In other words, we take a snapshot of Wikipedia, and model the associations between articles and their contributors based on the contribution from each user to the current revision of each article. By doing so, we intend to avoid the bias towards long-lived entities over newly created entities, which might be caused by the *chronological* approach [2]. In our previous research [17], we developed the Basic and PeerReview models based on the mutual dependency between article quality and contributor authority. This paper further extends this idea by introducing the review probability for each word as a contributor edits an article.

3. QUALITY MEASUREMENT MODELS

In this section, we introduce our article quality measurement models, namely Basic, PeerReview and ProbReview.

3.1 Notations

We first introduce the notations to be used throughout our discussion. Assuming there are N *article* entries in Wikipedia, we denote an article as a_i for $1 \leq i \leq N$; and, from the edit histories of all a_i's, we could identify M *users*, we denote a user as u_j for $1 \leq j \leq M$. The users' contribution to articles can be categorized into two types:

- **Authorship**: This involves the part of the content that originates from the user;

- **Reviewership**: This involves the part of the content that does not originate from the user, but, is reviewed by the user as he/she submits a newer revision of the article and keeps the reviewed content unchanged in the his/her revision.

Our basic unit of article content is a *word*. A word in an article a_i is denoted by w_{ik}. The relationship between users and a word can be denoted by:

- $w_{ik} \overset{A}{\leftarrow} u_j$, word w_{ik} is authored by user u_j; each word has exactly one author;

- $w_{ik} \overset{R}{\leftarrow} u_j$, word w_{ik} is reviewed by user u_j; a word could have zero or multiple reviewers.

The author and reviewer(s) of each word is identified by comparing the revisions of the article containing the word. Our quality measurement models therefore seek to compute:

- Q_i, the quality of each article a_i;

- A_j, the authority of each user u_j, as a by-product of computing article quality.

3.2 Basic

Our Basic model is designed based on the principle that "the higher authority are the authors, the better quality is the article." This principle measures the quality of an article by the aggregation of authorities from all its authors. However, an author's authority would then depend on the quality of articles the author has authored. Therefore, the two quantities, i.e., article quality and author authority, reinforce each other. Our Basic model is defined by Equations 1 and 2:

$$Q_i = \sum_j c_{ij} \times A_j \tag{1}$$

$$A_j = \sum_i c_{ij} \times Q_i \tag{2}$$

where c_{ij} denotes the amount of words u_j authored in a_i. Formally, $c_{ij} = \left|\{w_{ik} | w_{ik} \overset{A}{\leftarrow} u_j\}\right|$. By weighing the authority (or quality) values by c_{ij} when summing them up, Basic considers precisely the amount of contribution from the authors.

Similar to the link analysis in Web search [7, 11, 12, 14, 15, 22], the reinforcing quantities Q_i and A_j can be computed iteratively to derive their converged values. We discuss the computational issues of Basic together with our other models and the convergence properties of an iterative implementation in Section 3.6.

3.3 PeerReview

In Wikipedia, peer review on articles is ever on-going. An article may undergo a series of edits by contributors. When editing an article, the contributor would have first reviewed the prior content of the article, and then makes his/her own modifications. Content that survives through the new edit indicates the approval from the current contributor. If the content is approved by high authority reviewers, it is expected to be of better quality, even though the original authors of the reviewed content might be of low authority. Observing this peer reviewing process, we would like to incorporate reviewer authority into the definition of article quality. An intuitive way is to aggregate all reviewer authorities into content quality. Equations 3 and 4 give our

[6] http://en.wikipedia.org/wiki/Wikipedia:Featured_article

[7] http://en.wikipedia.org/wiki/Wikipedia:Clean_up

definition to PeerReview model:

$$q_{ik} = \sum_{w_{ik} \overset{A}{\leftarrow} u_j \cup w_{ik} \overset{R}{\leftarrow} u_j} A_j \tag{3}$$

$$A_j = \sum_{w_{ik} \overset{A}{\leftarrow} u_j \cup w_{ik} \overset{R}{\leftarrow} u_j} q_{ik} \tag{4}$$

In such definition of quality measurement model, each article is considered as a 'bag of words', formally $a_i = \{w_{ik}\}$. Hence, quality of the article sums up all word qualities, i.e., $Q_i = \sum_k q_{ik}$. It is worth noting that PeerReview model considers a user as the true reviewer of a word as long as:

1. The user has created a revision of the article that contains the word; and
2. In this revision, the word is taken from the article's previous revision and it remains unchanged by the user.

Therefore, whenever the relationship $w_{ik} \overset{R}{\leftarrow} u_j$ can be established between w_{ik} and u_j, the quality of word w_{ik} would take full credit from the authority of its reviewer u_j, and vice versa. Also, we consider each reviewer as important as the word's author, and there is no difference in weighing their contributions to derive word quality. Similarly, the reviewed words are credited to reviewer's authority in exactly the same way as the authored words. The intuition is based on the observation that "good contributors not only author but also review a considerable amount of good quality content".

3.4 ProbReview

PeerReview's assumption that each user, who edits the article content, would review the entire article prior to his/her edit is not always true. Consider these scenarios:

1. Certain statistics entries in an article were missing in an earlier revision. Some time later, a user looked up these missing entries from an external source, and submitted a newer revision with the missing statistics filled up.
2. Some Wikipedia users volunteered themselves for formatting and tweaking articles instead of actually adding more content. They could proficiently apply a template or re-organize the paragraphs, improving their visual appearance.

In both the above cases and many more, we are not absolutely certain that a reviewer will go through all other parts of an article when he or she edit only one portion of the article. In ProbReview, we therefore associate the relationship $w_{ik} \overset{R}{\leftarrow} u_j$ with a review probability $Prob(w_{ik} \overset{R}{\leftarrow} u_j)$. Recall that, the essence of identifying the reviewer(s) of a word is to imply their approval so as to apply their authorities to the word. ProbReview therefore modifies the definition of PeerReview model to capture the partial reviewership of each word.

We define ProbReview model in Equations 5 and 6:

$$q_{ik} = \sum_j f(w_{ik}, u_j) A_j \tag{5}$$

$$A_j = \sum_{i,k} f(w_{ik}, u_j) q_{ik} \tag{6}$$

where,

$$f(w_{ik}, u_j) = \begin{cases} 1 & \text{if } w_{ik} \overset{A}{\leftarrow} u_j \\ Prob(w_{ik} \overset{R}{\leftarrow} u_j) & \text{otherwise} \end{cases} \tag{7}$$

The function $Prob(w_{ik} \overset{R}{\leftarrow} u_j)$ is defined to return 0 in the following special cases:

1. $Prob(w_{ik} \overset{R}{\leftarrow} u_j) = 0$, when user u_j has never updated the content of article a_i;
2. $Prob(w_{ik} \overset{R}{\leftarrow} u_j) = 0$, when the word w_{ik} has never appeared in user u_j's revision(s) of article a_i.

To elaborate these ideas more precisely, we introduce the notation of timestamp, t_p. We denote the pth revision of an article a_i by $a_i^{t_p}$ where t_p denotes the time when the revision is created. Formally, $t_p < t_q$ for $p < q$. Our previous notation a_i corresponds to the latest revision of article i, i.e., $a_i \equiv a_i^{t_{\max_p}}$. If user u_j contributes the revision $a_i^{t_p}$, we denote it by $\mathcal{C}(a_i^{t_p}) = u_j$. Hence, user u_j has updated article a_i, if $\exists a_i^{t_p}$, such that $\mathcal{C}(a_i^{t_p}) = u_j$. Similarly, word w_{ik} has existed in user u_j's revision, if $\exists a_i^{t_p}$, such that $\mathcal{C}(a_i^{t_p}) = u_j \wedge w_{ik} \in a_i^{t_p}$.

Our next task is to determine $Prob(w_{ik} \overset{R}{\leftarrow} u_j)$ in cases where the value is non-zero. Intuitively, when a user authors some content in an article, other parts of the article that are closer to these authored content are more likely to be read. In other words, if there is a word w_{il} such that is authored by u_j, formally $w_{il} \overset{A}{\leftarrow} u_j$, then $Prob(w_{ik} \overset{R}{\leftarrow} u_j)$ should be larger if w_{ik} is close to w_{il}; it should decrease when w_{ik} is further away from w_{il}. Therefore, $Prob(w_{ik} \overset{R}{\leftarrow} u_j)$ can be modeled as a monotonically decaying function of d_{kl}, the distance between w_{ik} and w_{il}. When there are more than one such w_{il} in article a_i, $Prob(w_{ik} \overset{R}{\leftarrow} u_j)$ should take the maximum probability derived from all possible w_{il}. To summarize the above, we define $Prob(w_{ik} \overset{R}{\leftarrow} u_j)$ in the table shown below:

$\mathbf{Prob(w_{ik} \overset{R}{\leftarrow} u_j)}$	**Condition**
0	$\nexists {a_i}^{t_m}, \mathcal{C}({a_i}^{t_m}) = u_j$
0	$\nexists {a_i}^{t_m}, w_{ik} \in {a_i}^{t_m} \wedge \mathcal{C}({a_i}^{t_m}) = u_j$
$\max_l S(d_{kl})$	$\exists {a_i}^{t_m}, \mathcal{C}({a_i}^{t_m}) = u_j \wedge w_{ik} \in {a_i}^{t_m}$ and $\exists l, w_{il} \overset{A}{\leftarrow} u_j \wedge l \neq k$
0	otherwise

where, $S(d_{kl})$ is the review probability decaying scheme, as a function of the word distance d_{kl} from w_{ik} to w_{il}.

There are several candidate decaying schemes that can be used:

$$S^1(d_{kl}) = \frac{1}{|d_{kl}|} \tag{8}$$

$$S^2(d_{kl}) = \frac{1}{\max(|d_{kl}| - \alpha, 0) + 1} \tag{9}$$

$$S^3(d_{kl}) = \frac{1}{\sqrt{\max(|d_{kl}| - \alpha, 0) + 1}} \tag{10}$$

Figure 1 depicts these three candidate schemes. $S^1(d_{kl})$ decreases $Prob(w_{ik} \overset{R}{\leftarrow} u_j)$ at the word level. Even if w_{ik} and w_{il} are few words away, the chances that w_{ik} is reviewed by the author of w_{il} will drop to no more than half quickly.

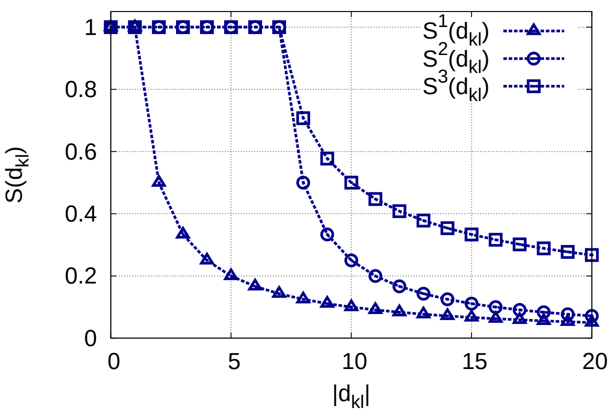


Figure 1: Review Probability Decaying Schemes

Whereas $S^2(d_{kl})$ and $S^3(d_{kl})$ emulate the sentence concept, by keeping $Prob(w_{ik} \overset{R}{\leftarrow} u_j)$ constant within α number of words before and after w_il. Words that belong to the same sentence as of w_{il} are most likely to be reviewed. While, for words that fall beyond the boundary of the sentence to which w_{il} belongs, their chances of being reviewed would then decrease as a function of the distance to the boundary. Parameter α can be regarded as the estimated average distance from w_{il} to the boundary of the sentence it belongs to. As shown in Figure 1, $Prob(w_{ik} \overset{R}{\leftarrow} u_j)$ drops quickly as d_{kl} increases beyond α. Therefore, $S^3(d_{kl})$ was intended to improve $S^2(d_{kl})$ by smoothing the decaying rate.

The estimated average sentence length, i.e., α in Equations 9 and 10, is the only parameter to tune in our refined ProbReview model.

3.5 Naïve

In contrast with our authority-based quality measurement models, a naïve way of judging article is based on length alone. The longer is the article, the better quality it is expected to carry. Equation 11 defines the Naïve model, which will be used as the baseline for performance evaluation.

$$Q_i = \sum_j c_{ij} \tag{11}$$

3.6 Computations

Basic, PeerReview and ProbReview are models that involve mutual dependency between quality and authority quantities in different forms. They can be implemented using iterative computation over a set of equations. This iterative computation process includes the following steps:

1. **initialize** all quality and authority values uniformly[8];
2. for each iteration:
   - use the authority values from the previous iteration to **compute qualities**;
   - use these quality values to **compute authorities**;
   - **normalize** all authority and quality values by L_1 norm;
3. repeat step 2 until the authority and quality values converge.

The convergence of such computation has been intensively addressed in [10, 15, 25]. If we represent quality values in vector $\vec{Q}$[9], of dimension D; all authority values in vector $\vec{A}$, of dimension M; and all interaction data between quality and authority in an adjacency matrix $\mathbf{M_d}$ of dimension $D \times M$. Given the condition that $\mathbf{M_d}$ is diagonalizable and has a unique largest eigenvalue[10] [10], the resulting quality vector $\vec{Q}$ would converge to the eigenvector of $\mathbf{M_d}$ corresponding to its largest eigenvalue, and authority vector $\vec{A}$ would converge to the corresponding eigenvector of $\mathbf{M_d}^T$, almost independently of the initial values used.

Besides absolute convergence, other terminating conditions commonly adopted in practice [11, 12, 14, 25] are:

- when the difference in all value changes from the previous iteration to the next is sufficiently small; or,
- when a predefined maximum number of iterations has been conducted.

Because of L_1 normalization, the resulting quality(authority) values are in the range of $[0, 1]$. These values preserve the relative ratios within article quality and contributor authority. The normalized quality score values that preserves the relative ratio among them are good enough for us to rank the articles.

3.7 Summary

To summarize, our Basic model is designed based on the mutual dependency between article quality and contributor authority. Only authors of the article are considered in this model. Our PeerReview model is built on top of Basic by introducing the role of word reviewers. Reviewer authority is treated equally as that of authors. Finally, our ProbReview model refines PeerReview by emulating the partial reviewership of each contributor. Words that are farther away from those that originate from a contributor are assumed less likely to have been reviewed by the contributor.

4. EXPERIMENTS

Our objective in conducting the experiments is two-fold. Firstly, we want to evaluate and compare the effectiveness of our proposed quality measurement models. Secondly, by varying some parameters and by incorporating length features of articles, we study the behavior of the proposed models in more details. In the following, we describe the dataset used, our proposed performance metric, and our experimental results.

[8] Random non-zero initialization is another option. It has been proven [10] that the final converged quality and authority values would be independent of the values used in initialization.

[9] $\vec{Q}$ entries may refer to article quality Q_i's as in Basic, or they may refer to word quality q_{ik}'s in PeerReview and ProbReview.

[10] For cases when $\mathbf{M_d}$ consists of multiple diagonalizable submatrices, where each can be linearly separated from one another, our intuitive solution is to decompose $\mathbf{M_d}$ and solve the equations for each sub-matrix separately. However, such case was not encountered in our datasets so far. Further investigation on these cases is scheduled in our future work.

Table 1: Distribution of Labeled Articles

Class	**FA**	**A**	**GA**	**B**	**Start**	**Stub**	Total
# articles	14	20	11	155	30	0	230
%	5.8	8.3	4.5	64.0	12.4	-	95.0
$s(p)$	4	3	2	1	0	-	-

Table 2: Dataset Statistics

Count		min	max	avg
# authors	per article	60	1058	227.6
# articles	per author	1	194	1.7
# words	per article	945	11,979	3,881.1
	per author	1	11,435	28.2
	per contribution	1	3,862	17.0
	per reviewer	0	834,572	2,437.43
# reviewers	per article	90	2,087	406.1
# articles	per reviewer	0	234	2.9

4.1 Dataset

We chose a set of 242 country articles in Wikipedia for our experiments. These article titles were obtained from the page "List of countries"[11] in Wikipedia. The main reason for choosing this set of articles was because the majority of them have been assigned class labels according to Wikipedia Editorial Team's quality grading scheme[12]. These manual labels were regarded as the ground truth in our model evaluation.

The label 'FA' stands for *Featured Articles*, which represent the best works in Wikipedia. 'A'-class articles provide a complete treatment to their subjects. 'GA' stands for *Good Articles*, which must meet the criteria of "well written, stable, accurate, referenced, have a neutral point of view, and show relevant illustrations with an appropriate copyright". 'B'-class is the next best. 'Start'-class articles are new articles being constructed, and 'Stub'[13]-class has the lowest quality status. We present them in decreasing quality in Table 1, i.e., FA ≥ A ≥ GA ≥ B ≥ Start ≥ Stub.

We crawled the latest revisions of the 242 articles dated 5th November, 2006, together with all past edit histories using MediaWiki's query API[14]. We also extracted the class labels of the articles from each article's talk page, dated 5th November, 2006. Table 1 also summarizes the class label distribution statistics of this set of articles. Note that none of the articles in this collection was labeled Stub. There were 12 articles left unlabeled, which represented less than 5% of the article collection. And, the majority were of B-class, which occupied 64% of the article collection.

4.2 Data Cleansing and Preprocessing

Article quality is measured on the latest revision of each article. The author and reviewer(s) for each word instance are extracted from the article edit history.

The edit history of articles consists of not only revisions submitted by human users, but also revisions created by robots (or 'bots' for short), which are automatic processes that interact with Wikipedia articles. The functions of editing bots are mainly: automatic importing, spell checking, wikifying, anti-vandalism and ban enforcement[15]. Since bot-created revisions generally does not carry content contribution, and revisions created by them do affect article quality measurement, we therefore decided to filter out revisions created by bots.

Besides bot revisions, we also removed a series of consecutive revisions from the same user by keeping only the last revision of the series. It was observed users had the habit of saving intermediate revisions to avoid loss of work due to unexpected hardware, software or network errors. By removing these intermediate revisions, we reduced the computation time without losing necessary interaction data.

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

As a result of data preprocessing, we identified 103,067 unique non-bot users from this article collection; 33,249 of them had authored at least one word in the latest revisions; and only 29.8% (i.e., 9,896) of these authors were registered users. Table 2 summarizes the statistics of our preprocessed dataset.

4.3 Evaluation Metrics

We adopt two evaluation metrics. The first metric is called **Normalized Discounted Cumulative Gain at top k** ('NDCG@k' for short) to evaluate the accuracy of the article ranking produced by a given quality measurement model. NDCG was first defined as an IR evaluation metric by Jarvelin et al [13] to consider the degree of relevance in retrieved results. More relevant results retrieved at top positions in the rank would accumulate higher score to the top k gain. This metric was chosen because it is particularly suited for ranked articles that have multiple levels of assessment, corresponding to the FA ≥ A ≥ GA ≥ B ≥ Start class labels.

$$NDCG = \frac{1}{Z} \sum_{p=1}^{k} \frac{2^{s(p)} - 1}{\log(1+p)} \tag{12}$$

As shown in Equation 12, NDCG@k is computed by summing up the gains from position $p = 1$ to $p = k$ in the ranking results. Given rank position p, $s(p)$ is an integer representing the amount of reward given to the article at position p. In our case, $s(p) = 4$ when the p-th ranked article has a FA label, $s(p) = 3$ for A-labeled article, and so on and so forth. We summarize $s(p)$ values and their corresponding article label at position p in the third row of Table 1. Note that Start-labeled or unlabeled article at position p does not contribute to the cumulative gain.

The term Z is a normalization factor derived from a perfect ranking of top k articles so that it would yield a NDCG

[11] `http://en.wikipedia.org/wiki/List_of_countries`

[12] `http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment`

[13] A Wikipedia stub is a very short article in need of expansion. See page `http://en.wikipedia.org/wiki/Wikipedia:Stub`

[14] `http://en.wikipedia.org/w/api.php`

[15] `http://en.wikipedia.org/wiki/Wikipedia:Types_of_bots`

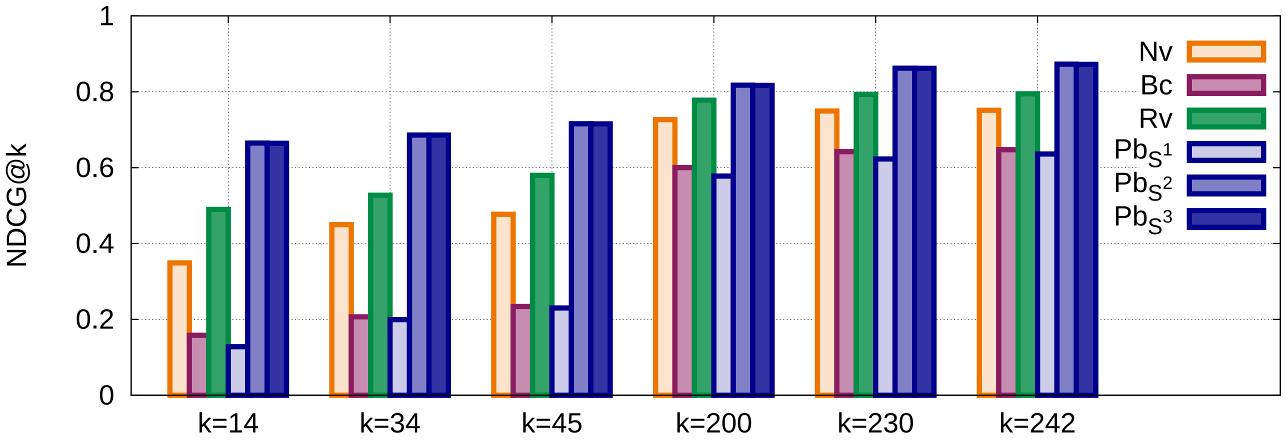


Figure 2: NDCG@k Performance, with $\alpha = 7$. (Nv: Naïve; Bc: Basic; Rv: PeerReview; Pb: ProbReview; and S^1, S^2 and S^3 corresponds to $S^1(d_{kl})$, $S^2(d_{kl})$ and $S^3(d_{kl})$ respectively)

of 1 [3]. Intuitively, the perfect ranking should place all FA-class articles before all A-class articles, followed by all GA-class articles and so on; finally, all Start-class articles should be place at the bottom end of the ranking result.

Our second metric is the Spearman's rank correlation coefficient. Spearman's rank correlation coefficient is a well-known metric for comparing the agreement between two rankings on the same set of objects.

Different from NDCG that considers articles belonging to the same quality class equally, Spearman's rank correlation coefficient preserves the one-to-one correspondence between articles in the two rankings. Another difference in these two metrics is in their value ranges. NDCG@k metric has value in the range of [0,1]. NDCG = 1 indicates perfect performance, and NDCG = 0 indicates worst performance. Whereas, Spearman's rank correlation coefficient has value in the range of [-1,1]. When two rankings on the same set of objects give a coefficient of 1, this means the two rankings are perfectly matched on the ordering of these objects; when they give a coefficient of -1, the two rankings are in exact reverse order; the median value 0 indicate total random correlation on these two rankings.

4.4 Results

4.4.1 NDCG@k

In this set of results, we compare NDCG@k among Naïve, Basic, PeerReview and ProbReview model with $S^1(d_{kl})$, $S^2(d_{kl})$ and $S^3(d_{kl})$ schemes defined in Equations 8, 9 and 10 respectively. We took six k values at 14, 34, 45, 200, 230 and 242 respectively. These k values were derived from our dataset statistics shown in Table 1. k was incremented each time by the total number of articles in the next best quality class.

For the ProbReview models, we used $\alpha = 7$. The choice of $\alpha = 7$ was based on the past empirical studies that said: (i) there are 5 to 35 words in one sentence for text documents [9]; and (ii) the span of immediate memory which imposed limitations on the amount of information that people were able to receive, process, and remember is "the magic number seven" [19]; and (iii) our brief calculation indicates an average of 6.04 words per sentence, with standard deviation of 5.85, for this set of Wikipedia articles after stop words removal. These studies suggest that $\alpha = 7$ is a reasonable value for our experiments. The NDCG@k values for Naïve, Basic, PeerReview, and ProbReview models are depicted in Figure 2.

It is clear from Figure 2 that our PeerReview and ProbReview models with $S^2(d_{kl})$ and $S^3(d_{kl})$ always outperform the baseline model Naïve for all k values. Especially, when k is small, the performance gaps between these three models and the Naïve model are seen more significant. On the down side, the Basic and ProbReview with $S^1(d_{kl})$ models do not give superior performance than Naïve.

When $k = 14$, the PeerReview and ProbReview models with $S^2(d_{kl})$ and $S^3(d_{kl})$ returned 5, 6 and 6 FA-class articles respectively in the top 14 ranked articles produced by their quality rankings. The Naïve model, on the other hand, only returned 3 FA-class articles in the top 14 ranked articles. Note that there are altogether 14 FA-class articles in our collection.

As k take larger values, not only the cumulative gain gets larger, but also the normalization factor grows. When $k = 242$, which includes all articles in our dataset, the resulting NDCG is an indicator of how well the overall ordering of all articles in our collection matched with the human manual assessment. From the right most block in Figure 2, it is clear that our authority-based PeerReview and ProbReview with appropriate schemes give promising performance in terms of overall article ranking.

The poor performance of $S^1(d_{kl})$ in ProbReview model could be caused by the drastic drops in $S(d_{kl})$ with increasing word distance d_{kl}. We suspected this reviewing behavior was presumably rare in practice. As a result, $S^1(d_{kl})$ do not perform as well as the other two schemes that incorporate a non-zero α.

Not surprisingly, the performance of Basic is comparable to that of ProbReview with $S^1(d_{kl})$. By the definition of $S^1(d_{kl})$ in ProbReview model, if we discard all review probabilities for $d_{kl} \neq 0$, ProbReview with $S^1(d_{kl})$ would degenerate into Basic, which considers only authority of word authors. However, because of the non-zero tail at $d_{kl} \neq 0$, $S^1(d_{kl})$'s performance deviates a little from that of Basic.

Table 3: Spearman's Rank Correlation Coefficients, with $\alpha = 7$

Model	Nv	Bc	Rv	$\mathbf{Pb}_{S^1}$	$\mathbf{Pb}_{S^2}$	$\mathbf{Pb}_{S^3}$
Nv	1.000	0.293	0.870	0.022	0.279	0.289
Bc	-	1.000	0.377	0.046	0.201	0.206
Rv	-	-	1.000	0.032	0.367	0.382
$\mathbf{Pb}_{S^1}$	-	-	-	1.000	0.504	0.506
$\mathbf{Pb}_{S^2}$	-	-	-	-	1.000	**0.998**
$\mathbf{Pb}_{S^3}$	-	-	-	-	-	1.000

Table 4: Variance in NDCG@k vs Parameter α

%	$\mathbf{Pb}_{S^2}$			$\mathbf{Pb}_{S^3}$		
	α=10	α=20	α=30	α=10	α=20	α=30
$k = 14$	0.000	0.000	0.000	0.000	0.000	0.000
$k = 34$	0.000	0.020	-0.008	0.000	0.019	0.024
$k = 45$	0.000	0.019	-0.007	0.000	0.008	0.004
$k = 200$	0.001	0.003	-0.020	-0.001	0.006	-0.004
$k = 230$	0.001	0.004	-0.018	-0.001	0.007	-0.004
$k = 242$	0.001	0.004	-0.018	-0.001	0.006	-0.004

On the whole, this set of results suggests that by carefully considering the authority of reviewers together with the authority of the actual authors (i.e., PeerReview and ProbReview with S^2 and S^3) in collaborative editing, we are able to achieve better quality ranking of articles than naïvely judging articles by length.

4.4.2 Spearman's Rank Correlation

Table 3 shows the Spearman's rank correlation coefficients between pairs of article quality rankings produced by various models. Same as in Figure 2, parameter α was set to 7 for $S^2(d_{kl})$ and $S^3(d_{kl})$ in ProbReview.

The highest rank correlation is observed between $S^2(d_{kl})$ and $S^3(d_{kl})$ of ProbReview. This observation agrees with the comparable performance between these two schemes in terms of NDCG. Recall our definition of these two review probability decaying schemes in Figure 1, $S^3(d_{kl})$ is intended to improve $S^2(d_{kl})$ by smoothing the decaying rate at d_{kl}s beyond the sentence boundary of w_{il}. As shown in Table 3, the high correlation between these two schemes suggests that these two decaying schemes did not alter the final ranking significantly in this experimental setting.

The lowest rank correlation is observed in the column of ProbReview with $S^1(d_{kl})$. Because of the assumption, that review probability decreases quickly with small d_{kl}, is presumably rare, $S^1(d_{kl})$ did not produce article quality ranking compare with Naïve, Basic and PeerReview models. However, because of the inherent formulation of ProbReview models, $S^1(d_{kl})$ is correlated with $S^2(d_{kl})$ and $S^3(d_{kl})$ to some extent.

4.4.3 Varying Parameter α

Other than using ProbReview model with $\alpha = 7$, we also explored the options of $\alpha = 10$, $\alpha = 20$ and $\alpha = 30$, to extend the boundary in which the review probability remains 1. In this case, we are interested in the way NDCG changes in the resulting article rankings with different α settings. We summarize the percentage of difference in NDCG as compared with that using $\alpha = 7$ in Table 4.

In general, when α grows, the variance in NDCG is observed to be small for all k. The same observation, that larger α does not alter the ranking of top 14 articles, holds for both $S^2(d_{kl})$ and $S^3(d_{kl})$ of ProbReview. Both schemes seem to favor $\alpha = 20$, as shown by the rise of positive variance. Larger α with smoother review probability decaying scheme in ProbReview shows smaller reduction in NDCG in Table 4. Theoretically, when $\alpha = \infty$, ProbReview with both $S^2(d_{kl})$ and $S^3(d_{kl})$ will degenerate into PeerReview.

Table 5: Average Words per Article for each Class

Class	FA	A	GA	B	Start
avg	5986.4	5810.4	5338.9	3748.7	2233.7
std dev	1682.7	2582.2	1431.3	1812.9	1327.6

4.4.4 Incorporating Article Length

Interestingly, while Naïve did not produce the best performance (see Figure 2), it performed better than Basic and ProbReview with $S^1(d_{kl})$. We therefore suspect a correlation between article length and article quality. Table 5 summarizes the average word count (after stop words removal) in articles for each quality class in our collection.

Table 5 shows that higher quality class corresponded to larger average article length. The three classes, FA, A and GA, were of comparable average article length, i.e., in the range of $5,330$ to $5,990$. There were, however, significant gaps in average article length between these three classes and class B articles, as well as between class B articles and class Start articles. It is noted that, in Wikipedia, FA- and GA-class articles require to be nominated as well as peer-reviewed before their status can be promoted. This empirical observation suggests that article length does play an important part in judging article quality in Wikipedia.

In this section, we therefore study how article length can possibly improve the performance of our proposed quality measurement models. The intuition is to bring good features together yet letting each play a part in final quality ranking. We essentially combine the Basic, PeerReview and ProbReview models linearly with the Naïve model as shown in Equation 13. This leads to the hybrid versions of the proposed models.

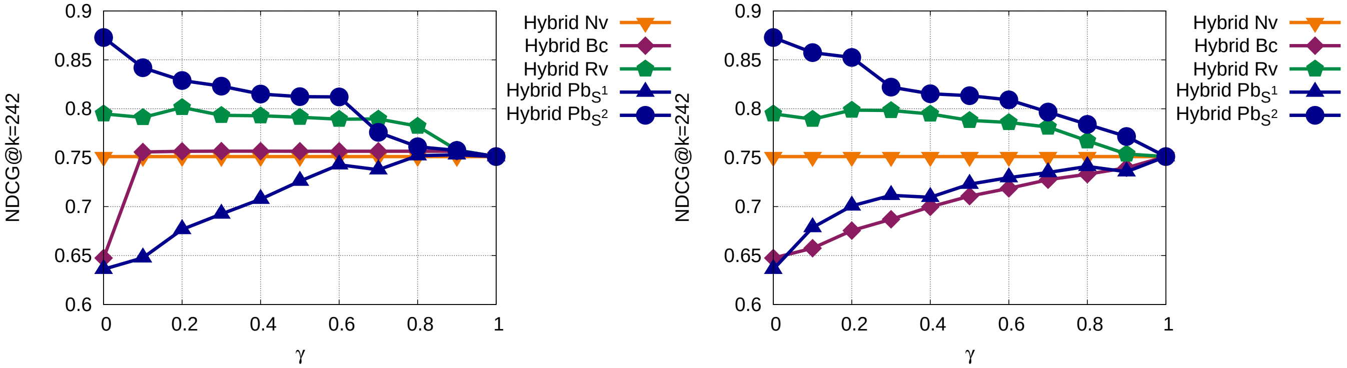
$$\tilde{g}(a_i) = \gamma \times g_0(a_i) + (1 - \gamma) \times g(a_i) \qquad (13)$$

In this equation, $\tilde{g}(a_i)$ denotes the combined quality measure for article a_i, while $g_0(a_i)$ and $g(a_i)$ represent the original measures given by Naïve model and any one of the other quality measurement models respectively. We experimented combination by quality scores and quality rank positions, as shown in Figure'3(a) and 3(b) respectively. For the hybrid ProbReview, we chose to use $\alpha = 7$ and $S^2(d_{kl})$ only. $S^3(d_{kl})$ is not reported in this section because it gave very similar results as $S^2(d_{kl})$.

Figures 3(a) and 3(b) depict the NDCG@k=242[16] for hybrid models by combining their computed quality scores and quality ranks respectively. We varied the γ values from 0 to 1 in 0.1 intervals. On the whole, with different γ values, the performance of proposed models remain largely unchanged in their hybrid versions, except for the following cases.

Hybrid Basic using quality score combination improves over the original Basic significantly such that it even out-

[16] 242 is the total number of articles in our country collection. NDCG@k=242 indicates a measure of overall ranking of articles in the assessment scale.



(a) Combine by Quality Scores

(b) Combine by Original Ranks

Figure 3: NDCG@k for Hybrid Models vs γ

Table 6: Contributor Statistics and Performance Measure on User Ranking

Contributor Statistics				Performance on User Ranking						
Category	**# users**	**avg # words**		**Score**	**k**	**NDCG@k**				
		authored	**reviewed**			**Bc**	**Rv**	**Pb$_{S^1}$**	**Pb$_{S^2}$**	**Pb$_{S^3}$**
Non-registered	23,353	16.4	1,252.3	0.0	33,249	0.881	**0.911**	0.878	0.872	0.873
Registered	9,816	56.2	5,235.3	1.0	9,896	0.408	**0.487**	0.384	0.372	0.376
WikiProject:Country Participants	80	821.6	22,540.4	2.0	80	0.222	**0.392**	0.293	0.255	0.259

performs Naïve slightly when $\gamma \geq 0.1$. This observation shows that, Basic model benefits from article length information. Nevertheless, the improvement does not make Basic much better than Naïve.

Hybrid PeerReview shows slightly better NDCG@k=242 at $\gamma = 0.2$. This improvement (i.e., 0.007 over the original PeerReview and 0.05 better than Naïve) is however not significant. When γ becomes larger, this improvement vanishes gradually. Hybrid ProbReview does not improve over the original ProbReview model for all the γ values. In fact, ProbReview model suffers from the article length information.

This set of results on the hybrid models shows that linear combination does not help improving PeerReview and ProbReview with $S^2(d_{kl})$ further. However, by incorporating length feature of articles, it improves Basic model over the baseline to some extent.

4.4.5 Contributor Authorities

In this section, we examine the derived authority of users by dividing them into 3 groups, namely "non-registered", "registered" and "WikiProject:Country participants". The "non-registered" users are identified by their IP addresses. The "registered" users are identified by their unique user names. "WikiProject:Country participants" are enthusiasts who volunteered themselves to constantly improving Wikipedia articles in the set of country subjects. Their user names were found on page "Wikipedia:WikiProject Countries/Participants"[17], as of 10 March 2007.

In Table 6, we show the overall statistics of these three user groups. Intuitively, we expected that, registered users should be more authoritative than non-registered users since the former's user names are known and their efforts would affect their reputation. The "WikiProject:Country participants" belong to an even more exclusive user group and are expected to be more authoritative. We computed NDCG@k for users rankings. Values of k were taken at 80, 9896 and 33249, corresponding to number of "WikiProject:Country participants", "WikiProject Country participants" cum registered users, and all users, respectively.

As Table 6 shows, PeerReview yields the best NDCG performance on user ranking for all three k. Interestingly, the best model for article ranking, i.e., ProbReview model with $S^2(d_{kl})$ and $S^3(d_{kl})$, does not perform well in user ranking. We believe it is due to voluntary participation in WikiProjects. Contributors of the high quality contribution may not volunteer themselves to participate in the "WikiProject: Countries" project. On the other hand, unregistered user could also give very high quality contributions [5]. Therefore, our previous assumption about different authority levels of contributors might be too general to represent the ground truth.

5. CONCLUSION

In this paper, we study models for automatically deriving Wikipedia article quality rankings based on the interaction data between articles and their contributors.

Our PeerReview model, which was first proposed in [17], had already shown promising performance over the baseline model Naïve. We further extended it to emulate the probability of article content being reviewed by each contributor. As shown in our experiments, the extended ProbReview models with review probability decaying schemes $S^2(d_{kl})$ and $S^3(d_{kl})$ were the best performers compared with all other models under the same setting. By observing that, user interaction data itself is not sufficient in judging article quality and article length appears to have some merits in identifying quality articles, we incorporated article length

[17] `http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Countries/Participants`

into article quality measurement. Our experimental results showed some performance improvement by Hybrid Basic and hybrid PeerReview models at $\gamma = 0.1$ and $\gamma = 0.2$ respectively. However, ProbReview models, did not benefit from article length.

Our evaluation and analysis in this paper have been focusing on article quality rankings for collaboratively authored content. However, assessing content quality and contributor authority are complementary steps in our data-driven approach. In the future work, we plan to investigate customized review behavior in co-editing, and to derive contributor expertise is the first step. Besides, we also plan to apply our proposed models on a much larger Wikipedia article set. Managing model scalability and integrating quality measurements with quality checking procedures in collaborative authoring are other steps forward in this research.

6. ACKNOWLEDGMENTS

This work was supported in part by A*STAR Public Sector R&D, Project Number 062 101 0031. We gratefully acknowledge the valuable comments and suggestions from our anonymous reviewers.

7. REFERENCES

[1] S. F. Adafre and M. de Rijke. Discovering missing links in Wikipedia. In *Proc. of LinkKDD'05*, pages 90–97, 2005.

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

[3] E. Agichtein, E. Brill, and S. Dumais. Improving Web search ranking by incoporating user behavior information. In *Proc. of SIGIR'06*, pages 19–26, 2006.

[4] R. B. Almeida, B. Mozafari, and J. Cho. On the evolution of Wikipedia. In *Proc. of ICWSM'07*, March 2007.

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

[6] T. Cross. Puppy smoothies: Improving the reliability of open, collaborative wikis, 2006. Retrieved online: `http://www.firstmonday.org/issues/issue11_9/cross/index.html`.

[7] C. Dwork, R. Kumar, and M. Naor. Rank aggregation methods for the Web. In *Proc. of WWW'01*, pages 613–622, 2001.

[8] J. Giles. Internet encyclopaedias go head to head, 2005. Published online: 14 December 2005 `http://www.nature.com/news/2005/051212/full/438900a.html`.

[9] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: Sentence selection and evaluation metrics. In *Proc. of SIGIR'99*, pages 121–128, 1999.

[10] G. H. Golub and C. F. V. Loan. *Matrix Computations*. Johns Hopkins University Press, 3rd edition, 1996.

[11] Z. Gyöngyi, P. Berkhin, H. Garcia-Molina, and J. Pedersen. Link spam detection based on mass estimation. In *Proc. of VLDB'06*, pages 439–450, 2006.

[12] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating Web spam with TrustRank. In *Proc. of VLDB'04*, pages 576–587, 2004.

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

[14] G. Jeh and J. Widom. Scaling personalized Web search. In *Proc. of WWW'03*, pages 271–279, May 2003.

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

[16] 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.

[17] 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, December 2006.

[18] Max Völkel and Markus Krötzsch and Denny Vrandečić and Heiko Haller and Rudi Studer. Semantic Wikipedia. In *Proc. of WWW'06*, pages 585–594, 2006.

[19] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Pychological Review*, 63:81–97, 1956.

[20] B. B. C. News. Wikipedia survives research test, 2005. Published online: 15 December 2005 `http://news.bbc.co.uk/2/hi/technology/4530930.stm`.

[21] A. Orlowski. Wikipedia founder admits to serious quality problems, 2005. Published online: 18 October 2005 `http://www.theregister.co.uk/2005/10/18/wikipedia_quality_problem`.

[22] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Technical report, November 1999.

[23] P. Schönhofen. Identifying document topics using the Wikipedia category network. In *Proc. of WI'06*, pages 456–462, 2006.

[24] M. Strube and S. P. Ponzetto. Wikirelate! computing semantic relatedness using Wikipedia. In *Proc. of AAAI'06*, pages 1419–1424, 2006.

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

[26] J. Voss. Measuring Wikipedia. In *Proc. of the 10th International Conference of the International Society for Scientometrics and Informatics*, pages 221–231, July 2005.

[27] J. Wales. Wikipedia sociographics, 2004. Retrieved online: `www.ccc.de/congress/2004/fahrplan/files/372-wikipedia-sociographics-slides.pdf`.

[28] Wikipedia. Replies to common objections, 2007. `http://en.wikipedia.org/wiki/Replies_to_common_objections` Accessed on April 2007.

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