On Ranking Controversies in Wikipedia: Models and Evaluation[*]

Ba-Quy Vuong, Ee-Peng Lim, Aixin Sun, Minh-Tam Le, Hady Wirawan Lauw, Kuiyu Chang
School of Computer Engineering
Nanyang Technological University
Singapore 639798
{vuon0001, aseplim, axsun, lemi0001, hady0002, askychang}@ntu.edu.sg

ABSTRACT

Wikipedia[1] is a very large and successful Web 2.0 example. As the number of Wikipedia articles and contributors grows at a very fast pace, there are also increasing disputes occurring among the contributors. Disputes often happen in articles with controversial content. They also occur frequently among contributors who are "aggressive" or controversial in their personalities. In this paper, we aim to identify controversial articles in Wikipedia. We propose three models, namely the *Basic* model and two *Controversy Rank* (*CR*) models. These models draw clues from collaboration and edit history instead of interpreting the actual articles or edited content. While the Basic model only considers the amount of disputes within an article, the two Controversy Rank models extend the former by considering the relationships between articles and contributors. We also derived enhanced versions of these models by considering the *age* of articles. Our experiments on a collection of 19,456 Wikipedia articles shows that the Controversy Rank models can more effectively determine controversial articles compared to the Basic and other baseline models.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: [Search Process]

General Terms

Algorithms, Experimentation

Keywords

Wikipedia, controversy rank, online dispute

1. INTRODUCTION

1.1 Motivation

With increasing popularity of Web 2.0 applications such as wikis, blogs and social tagging software, Web users now can easily edit, review and publish content collaboratively. One of the largest and most successful Web 2.0 examples is Wikipedia [14], a free online encyclopedia with approximately 7 millions articles in 251 languages, among which more than 1.9 millions are from the English Wikipedia[2]. Wikipedia has more than 4 millions registered users and is currently the 12th most popular website according to Alexa.com. The content of Wikipedia is highly regarded among the online communities [3, 6, 4, 12, 15]. According to a recent comparison conducted by *Nature*, the scientific entries in Wikipedia are of quality comparable to the those in the established Britannica encyclopedia [6].

As Wikipedia is growing very fast in both number and size [16], there is also a higher likelihood for disputes or controversies to occur among contributors. According to the Random House Unabridged Dictionary [5], *controversy* is defined as "a prolonged public dispute, debate, or contention; disputation concerning a matter of opinion". In this research, we are primarily concerned with controversial articles that attract disputes between contributors. An article is considered controversial if it has more potential for disputes to occur throughout its edit history. For example, "Holiest sites in Islam" is a very controversial Wikipedia article which has attracted much dispute among its contributors as they have difficulties agreeing on the list of holiest sites and especially their rank order. Several contributors even requested to change the article title or delete the article altogether.

In this research, we aim to identify those controversial articles, which is important for the following two reasons:

- Controversies appearing in Wikipedia articles are often a good reflection or documentation of the real world. Finding controversies in Wikipedia can therefore help the general public and scholars to understand the corresponding real world controversies better.
- It allows moderators and contributors to quickly identify highly controversial articles, thereby improving the effectiveness of the dispute resolution process by reducing the amount of effort searching for such articles.

However, determining controversial articles is a challenging task due to several reasons:

[*]The work was supported in part by A*Star Public Sector R&D, Project Number 062 101 0031.

[1] http://wikipedia.org

[2] http://en.wikipedia.org

WSDM'08, February 11–12, 2008, Palo Alto, California, USA.
Copyright 2008 ACM 978-1-59593-927-9/08/0002 ...$5.00.

- *Large number of articles:* With the huge number of articles in Wikipedia each having its own edit history [2], it is not feasible to manually identify controversial articles by analyzing every article and its edit history.

- *Diverse content among articles:* Wikipedia articles cover a wide range of topics. It is difficult to recruit experts from different fields to perform content analysis on the edited article content.

- *Evolving content:* Wikipedia is always growing and its article content changes continuously. It poses further challenges to monitor the degree of controversy, which may vary over time.

Wikipedia currently attempts to identify controversial articles by allowing users to manually assign controversial tags to articles. Once tagged, the article title will appear in some special pages such as *Wikipedia:List of controversial issues*[3] for further editing by Wikipedia users. For example, the "Holiest sites in Islam" article has at least 460 controversial tags assigned to its different revisions. Since not all articles will be manually examined for controversy, there could potentially be many untagged articles that contain a lot of disputes.

1.2 Overview of Our Approach

In this research, we aim to automate the process of identifying controversial articles. Instead of interpreting article or edit content, we draw clues from the interaction among contributors recorded in the edit histories of articles. First, we define a *dispute* between a pair of users in a given article as the number of words that they have deleted from each other in the article's edit history. Note that there are two basic edit operations in Wikipedia: adding and deleting words. Replacement of words can be seen as deleting old words and adding new ones.

We are only interested in delete operations as they indicate disagreement between contributors. The collection of articles with content deleted by contributors in Wikipedia can be represented as a bipartite graph as shown in Figure 1. The graph consists of a set of contributor ordered pairs (u_i, u_j) and a set of articles (r_k). The directed edges from contributor pairs to articles represent disputes. Each edge is also assigned an integer value d_{ijk} to represent the amount of dispute measured by the number of contributor u_j's words deleted by contributor u_i in article r_k.

In our proposed approach, we measure the controversy in an article by the amount of dispute occurring in articles and the degree of controversy in each dispute. In general, a dispute in an article is more controversial if it occurs between two contributors who are known to cause little controversy. Conversely, a dispute between two contributors in an article is more controversial if the dispute takes place in a less controversial article.

On the whole, we have made the following novel contributions:

- We develop models to identify article controversy. They are the *Basic model* and the *Controversial Rank models* (*CR models*). These models work based on the amount and controversy level of disputes, which are derived from the articles' edit histories.

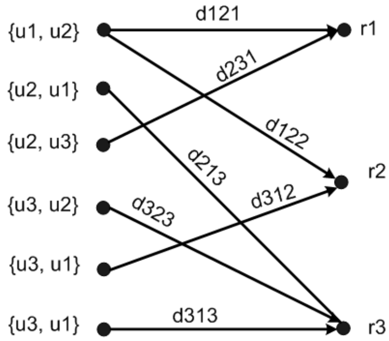


Figure 1: Disputes represented by a bipartite graph.

- We implement the different proposed models and evaluate them on a collection of more than 19,000 Wikipedia articles. In addition, we also introduce two baseline models for comparison. We conclude that our proposed CR models are able to rank controversial articles more accurately than other models. Some interesting cases will be discussed in greater detail.

- We further improve the CR models by incorporating the article and contributor *age* factors. The Age-aware CR models are designed to prevent frequent alterations in new articles from being mistaken as disputes since large number of edit operations are considered normal in these new articles.

1.3 Paper Outline

This paper is organized as follows. We examine the use of dispute tags and some baseline approaches to identify controversial Wikipedia articles in Section 2. Details of our proposed models are given in Section 3. We then describe our experimental setup together with results in Section 4. Section 5 presents the age-aware versions of the Controversy Rank models and some experimental results. We compare our work with other related projects in Section 6. Finally we conclude the paper in Section 7.

2. DISPUTE TAGS AND BASELINE MODELS

2.1 Dispute Tags

Articles in Wikipedia undergo changes made by contributors. A revision of an article refers to a version created and saved by a contributor [14]. Over time, articles can accumulate large number of revisions. If an article is observed to have disputes, its contributors can assign dispute tags to it flagging the article for attention. Dispute tags are tags defined by Wikipedia[4] to indicate different types of disputes occurring in articles. In this research, we are only interested in six types of dispute tags: "{{disputed}}", "{{totallydisputed}}", "{{controversial}}", "{{disputed-section}}", "{{totallydisputed-section}}"and "{{pov}}" (see Table 1). These tags are those that we consider relevant to the controversy of articles. Once an article is labeled with one of these tags, a message will appear on the top of the article

[3] `http://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues`

[4] `http://en.wikipedia.org/wiki/Wikipedia:Template_messages/Disputes`

Table 1: Dispute Tags Used

Tag	Meaning
{{disputed}}	The factual accuracy of this article is disputed.
{{totallydisputed}}	The neutrality and factual accuracy of this article are disputed.
{{controversial}}	This is a controversial topic, which may be under dispute.
{{disputed-section}}	Some section(s) has content whose accuracy or factual nature is in dispute.
{{totallydisputed-section}}	The neutrality and factual accuracy of some section are disputed.
{{pov}}	The neutrality of the article is disputed.

or article talk page and the article will be classified into the controversial category.

Naturally, articles with more dispute tags in their history are more controversial. To quantify the degree of article controversy, we define *Article Tag Count* (ATC_k) of an article r_k as the total number of tags appearing in all its revisions (see Equation 1).

$$ATC_k = \sum_{i=1}^{N_k} c_{ki} \qquad (1)$$

In this Equation, N_k and c_{ki} refer to the number of revisions of article r_k and the number of dispute tags found in the i^{th} revision of r_k respectively.

One should note that dispute tags only appear in a small number of articles. There are certainly other controversial articles that have not been tagged. Hence, the ground truth derived from the tagged controversial articles is not complete. Nevertheless, it still serves as a useful partial ground truth for performance evaluation.

2.2 Revision Count and Contributor Count

One simple way to identify degree of controversy in an article is to measure the number of revisions the article has undergone. Another equally simple way is to measure the number of contributors the article has. The former assumes that the more revisions the article has undergone, the more controversial it is. The latter applies a similar assumption on number of contributors.

To examine how closely the revision count and contributor count are related to article controversy, we analyse the statistics derived from a set of 19,456 articles from the *Religious Objects* category in Wikipedia. This subcollection is chosen because many topics (or articles) in this category are known to be controversial. There are altogether 71 articles assigned with dispute tags. More details about this dataset will be given in Section 4.1.

Figures 2(a) and (b) show the number of revisions of articles, and ATC values of articles, both sorted by the number of revisions (shown in log scale), respectively. It can be seen from Figure 2(a) that articles with large numbers of revisions are very few and most articles have very small numbers of revisions. As shown in Figure 2(b), the articles with large number of revisions are more likely to have large ATC values. We observe the same trend when similar figures are plotted based on numbers of contributors of articles.

We therefore define two baseline models based on the above observations. The **Revision Count model** uses the number of revisions to predict how controversial articles are. The **Contributor Count model** uses the number of contributors involved in an article as a measure of controversy.

Nevertheless, these simplistic approaches have some obvious drawbacks. They do not recognise disputes, which is the main factor contributing to an article's controversy. The Revision Count model can also be easily abused; one may edit the articles many times to increase its degree of controversy. The Contributor Count model may also wrongly classify heavily edited high quality articles to be controversial.

Tables 2 and 3 show the top 20 articles (together with their ATC values) returned by the Revision Count model and Contributor Count model respectively. As none of these articles are labeled with dispute tags, this clearly suggests that using number of revisions or number of contributors alone cannot detect highly controversial articles.

3. PROPOSED MODELS

3.1 Contribution Matrix

Before describing our models, we define the contribution of a contributor to an article as the number of words s/he has contributed to that article over its entire history. Similar to the bipartite graph representation of disputes (see Figure 1), contribution made by contributors to articles can also be represented by a bipartite graph as shown in Figure 3.

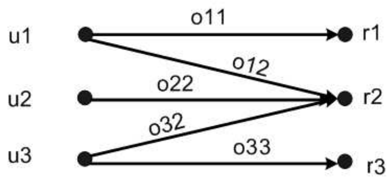


Figure 3: Contributions represented by a bipartite graph.

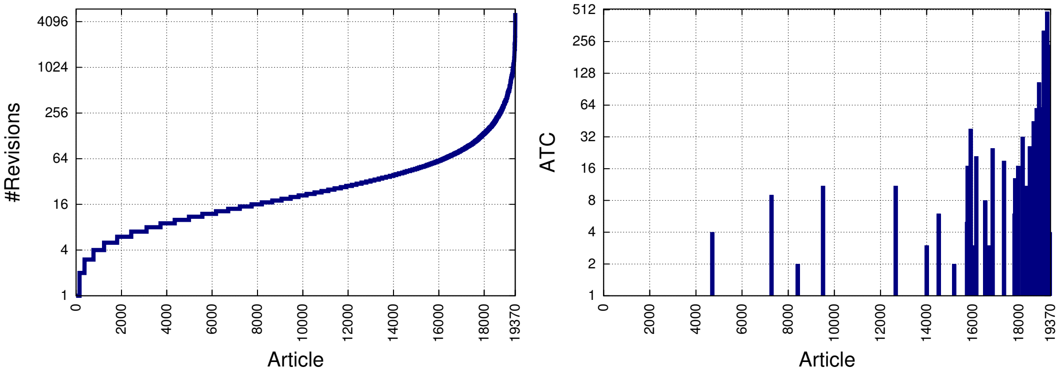
Vertices on the left of the graph represent contributors (u_i). Vertices on the right represent articles (r_k). The directed edge from contributor u_i to article r_k is assigned a weight o_{ik} representing the number of words contributed by u_i to r_k.

3.2 Basic Model

We now present our first proposed model for ranking controversial articles known as **Basic**. The Basic model measures the controversy of an article r_k, C_k^r, using Equation 2.

$$C_k^r = \frac{\sum_{i,j} d_{ijk}}{\sum_i o_{ik}} \qquad (2)$$

Recall that d_{ijk} is the number of words from contributor u_j to article r_k that are deleted by contributor u_i. We use d_{ijk} to capture the amount of dispute between u_i and u_j. This is because most disputes involve one contributor



(a)

(b)

Figure 2: #Revision and ATC Values of Articles Ordered by #Revision

Table 2: Top Revision Count 20 Articles

No	Article Name	#Rev	ATC	ATC Rank	No	Article Name	#Rev	ATC	ATC Rank
1	Podcast	5344	0	>71	11	Iain Lee	2384	0	>71
2	Emma Watson	4115	0	>71	12	Globe Theatre	2330	0	>71
3	Stephen Hawking	3682	0	>71	13	Grigori Rasputin	2211	0	>71
4	John McCain	3072	0	>71	14	John Howard	2200	0	>71
5	George Orwell	2948	0	>71	15	Keira Knightley	2177	0	>71
6	David Cameron	2692	0	>71	16	Salem witch trials	2176	0	>71
7	Dr.Seuss	2625	0	>71	17	Easter	2146	0	>71
8	James Madison	2623	0	>71	17	Constantine I	2111	0	>71
9	John Locke	2477	0	>71	19	Winston Churchill	2100	0	>71
10	Oscar Wilde	2432	0	>71	20	Rupert Murdoch	2003	0	>71

deleting the words added by another contributor. The denominator $\sum_i o_{ik}$ represents the total contribution received by article r_k.

Based on the same idea, we derive the Basic model for measuring contributor controversy. A *contributor* is said to be *controversial* if s/he is likely to be involved in disputes with others. In the Basic model, we determine contributor controversy using Equation 3.

$$C_i^u = \frac{\sum_{j,k}(d_{ijk} + d_{jik})}{\sum_{j,k} o_{jk} \times I(i,j,k) + \sum_k o_{ik}} \qquad (3)$$

In this equation, C_i^u is the controversy of contributor u_i. d_{ijk} (d_{jik}) is the number of words authored by u_j (u_i) in article r_k subsequently deleted by r_i (r_j). The numerator component $\sum_{j,k} d_{ijk}$ represents the amount of *active dispute* u_i has caused in articles that s/he has edited. The dispute is active as it is a result of deletion operations performed by u_i. The second numerator component $\sum_{j,k} d_{jik}$ represents the amount of *passive dispute* u_i has suffered in the articles s/he has edited. The dispute is passive as it is a result of deletion operations performed by others. The boolean function $I(i,j,k)$ indicates whether u_i has deleted any word from u_j in article r_k and is defined by Equation 4. The denominator $\sum_{j,k} o_{jk} \times I(i,j,k)$ and $\sum_k o_{ik}$ therefore represent the total contribution from users having disputes with u_i in those disputed articles, and the total contribution of u_i respectively.

$$I(i,j,k) = \begin{cases} 1 & \text{if } d_{ijk} > 0, \\ 0 & \text{if } d_{ijk} = 0. \end{cases} \qquad (4)$$

3.3 Controversy Rank Models (CR Models)

Throughout its history, an article may experience a number of disputes among its contributors depending on the topic it covers. In the **Controversy Rank** (CR) Models, we are interested in disputes caused by the nature of the articles, not by the “aggressiveness” of the contributors, i.e., the controversial contributors. Given an article, disputes involving controversial contributors are more likely caused by the contributors’ behavior, not the article topic. On the other hand, disputes involving non-controversial contributors should be rare. Such disputes are therefore strong evidence of article controversy. CR Models therefore derive the controversy score of an article by making reference to the controversy of its contributors.

To distinguish the different classes of disputes, CR models need to derive a controversy score for each contributor. Controversial contributors are generally expected to be involved in large number of disputes in different articles, even in the non-controversial articles. The non-controversial contributors, in contrast, are expected to avoid disputes. Hence, articles with disputes involving non-controversial contributors are likely to carry controversial content.

The above idea can be described using the following *Mutual Reinforcement Principle*:

Table 3: Top Contributor Count 20 Articles

No	Article Name	#Cont	*ATC*	*ATC* Rank	No	Article Name	#Cont	*ATC*	*ATC* Rank
1	Podcast	2146	0	>71	11	Easter	1041	0	>71
2	Stephen Hawking	1933	0	>71	12	Keira Knightley	1030	0	>71
3	Emma Watson	1619	0	>71	13	Rupert Murdoch	1028	0	>71
4	John McCain	1459	0	>71	14	Globe Theatre	974	0	>71
5	George Orwell	1342	0	>71	15	Winston Churchill	948	0	>71
6	Dr. Seuss	1186	0	>71	16	David Cameron	912	0	>71
7	John Locke	1174	0	>71	17	Salem witch trials	899	0	>71
8	Grigori Rasputin	1110	0	>71	18	Jefferson Davis	879	0	>71
9	Oscar Wilde	1093	0	>71	19	Constantine I	871	0	>71
10	James Madison	1085	0	>71	20	Robert Frost	848	0	>71

- *Article Controversy:* An article is more controversial if it contains more disputes among less controversial contributors.
- *Contributor Controversy:* A contributor is more controversial if s/he is engaged in more disputes in less controversial articles.

The CR models, based on the above mutual reinforcement principle, define the controversy scores of articles and contributors by Equations 5 and 6 respectively.

$$C_k^r = \frac{\sum_{i,j} agg[(1-C_i^u),(1-C_j^u)] \times d_{ijk}}{\sum_i o_{ik}} \quad (5)$$

$$C_i^u = \frac{\sum_{j,k}(1-C_k^r) \times (d_{ijk}+d_{jik})}{\sum_{j,k} o_{jk} \times I(i,j,k) + \sum_k o_{ik}} \quad (6)$$

In Equation 5, the article controversy score is measured by the sum of disputes in the article weighted by the aggregated inverse controversy of the pairs of contributors involved in the disputes. The aggregate function agg derives an aggregated inverse controversy score for two contributors u_i and u_j. The two arguments of the agg function represent the inverse controversy scores of u_i and u_j, which are computed as $(1-C_i^u)$ and $(1-C_j^u)$, respectively.

In Equation 6, the contributor controversy score is measured by the sum of both active and passive disputes in which contributor u_i is involved, weighted by the inverse controversy scores of the articles edited by u_i.

There are two options for the aggregation function agg, namely:

- *Average:*

  $$agg[(1-C_i^u),(1-C_j^u)] = \frac{1-C_i^u+1-C_j^u}{2} \quad (7)$$

  This function treats two inverse controversy values equally. If the two contributors in dispute are highly controversial, the aggregate function has a low value and vice versa. Dispute between two contributors with intermediate controversy results in an intermediate aggregate value. The CR model using average Agg function is called **Controversy Rank Average** (**CR Average**).

- *Product:*

  $$agg[(1-C_i^u),(1-C_j^u)] = (1-C_i^u) \times (1-C_j^u) \quad (8)$$

  In this option, disputes are only significant if they are between contributors with intermediate or low controversy. If they occur between a very high controversy user and a very low controversy user, the aggregated value will be close to zero. The CR model using product Agg function is called **Controversy Rank Product** (**CR Product**).

Equations 5 and 6 extend Equations 2 and 3 respectively by introducing two new factors: $agg((1-C_i^u),(1-C_j^u))$ and $(1-C_k^r)$. When every contributor has the same controversy score, and every article has the same controversy score, the CR Average (and CR Product) models will degenerate into the Basic model.

4. EXPERIMENTS AND RESULTS

4.1 Dataset

To evaluate the performance of our proposed models and compare them with other models, we conducted experiments on a category of articles from Wikipedia. We chose the *Religious Objects* category, a sub-category of *Religion*, because it is known to contain sensitive topics and several user tagged controversial articles. We downloaded a total of 19,456 articles from the category and sub-categories together with their edit histories on 12 June 2007. As some sub-categories included are less related to religion, our dataset also includes some articles from the non-religious domains. There are 174,338 distinct contributors in this dataset.

Next, we extracted all the contributors who have edited these articles and constructed the dispute bipartite graph (d_{ijk}'s) and contribution bipartitle graph information (o_{ik}'s) using the following algorithms:

- *Contribution Extraction:* We compare two successive revisions in the edit history of an article. Words in the new revision that do not exist in the old one are considered contributed by the contributor who creates the new revision. Words in the first revision of the article are considered to be contributed by the creator.
- *Dispute Extraction:* We also compare two successive revisions as in the above algorithm. Words in the old revision that do not exist in the new revision are considered deleted by the new revision's contributor. These words are then counted towards the dispute between the old and new contributors.

In applying the above two algorithms, we only considered successive revisions that were not created by the same user. It is common in Wikipedia that users usually save intermediate revisions to avoid work loss due to unexpected hardware,

Table 4: Dataset Statistics

Count		*Min*	*Max*	Avg	Std Dev
# contributors	per article	1	3190	39.69	99.47
# articles	per contributor	1	937	2.71	23.71
Contributions	per article	3	360,929	1324.21	9118.68
	per contributor	0	347,727	140.22	3092.58
Disputes	per article	0	359,165	902.82	8863.65
	per contributor	0	348,800	95.60	3888.33

Table 5: Distribution of ATC Values

ATC	>500	101-500	21-100	5-20	1-4	0	Total
# articles	0	6	19	21	25	19,385	19,456
%	0.0	0.031	0.098	0.108	0.128	99.635	100

software or network failures. Thus, we only considered the last revision in a succession of revisions made by the same user. All intermediate revisions were ignored. This practice not only reduced processing effort but also the running time of the CR models. In addition, we removed common stop words such as "a", "an", "the", etc. from the revisions.

As shown in Table 4, there are on average 39.69 contributors per article and each contributor contributes to on average 2.71 articles. These two numbers may vary a lot among articles and contributors. We also observe that each article has on average 1324.21 words and 902.82 words involved in disputes. On average, each contributor has contributed 140.22 words and is involved in 95.60 disputed words. It is possible for a contributor to have zero contribution when s/he only performs deletion(s).

We also calculated the ATC values of articles in the dataset. Among the 19,456 articles, only 71 have non-zero ATC values, as shown in Table 5.

4.2 Evaluation Metrics

To evaluate the effectiveness of the proposed models, we adopt two evaluation metrics. They are the *Precision-Recall-F1 at top k* (Precision-Recall-F1@k) and *Normalized Discounted Cumulative Gain at top k* (NDCG@k). As we only have a small set of labeled controversial articles (with $ATC \neq 0$), we are confined to applying these two evaluation metrics to the set of articles.

4.2.1 Precision, Recall and F1

Precision, recall and F1 are performance measures in Information Retrieval (IR) [8] that considers the number of relevant documents among top k documents. In our evaluation, relevant documents refer to the articles with non-zero ATC values.

$$\mathit{Precision@k} = \frac{\mathit{\#relevant\ articles\ in\ top\ k\ articles}}{k} \tag{9}$$

$$\mathit{Recall@k} = \frac{\mathit{\#relevant\ articles\ in\ top\ k\ articles}}{\mathit{\#relevant\ articles\ in\ the\ dataset}} \tag{10}$$

$$\mathit{F1@k} = \frac{2 \times \mathit{Precision@k} \times \mathit{Recall@k}}{\mathit{Precision@k} + \mathit{Recall@k}} \tag{11}$$

Note that the above precision, recall and F1 definitions do not concern matching the computed controversy rank order with the ground truth (ATC) rank order. We therefore introduce another performance metric known as NDCG.

4.2.2 NDCG

Normalized Discounted Cumulative Gain at top k (NDCG@k) is an evaluation metric in IR for measuring the rank accuracy of search results [8]. Unlike precision, recall and F1, NDCG recognizes the different levels of relevance and prefers rankings that follow the actual relevance order. This metric is particularly suited for ranking articles that have multiple levels of relevance. In our experiments, we use ATC values of articles to rank the controversial articles. The formula of NDCG is given in Equation 12.

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

In Equation 12, NDCG@k is computed as the sum of gains from position $p = 1$ to $p = k$ in the ranking results. $s(p)$ is the function that represents reward given to the article at position p. In our experiment, $s(p)$ is calculated based on the ATC value of the pth article: $s(p) = \log(ATC_p + 1)$. We use the log function to dampen the very large ATC values of highly controversial articles. Note that articles with zero ATC values do not contribute to the cumulative gain.

The term Z is a normalization factor derived from perfect ranking of the top k articles so that it would yield a maximum NDCG value of 1. The perfect ranking is one that places all articles in decreasing ATC order.

4.3 Overall Results

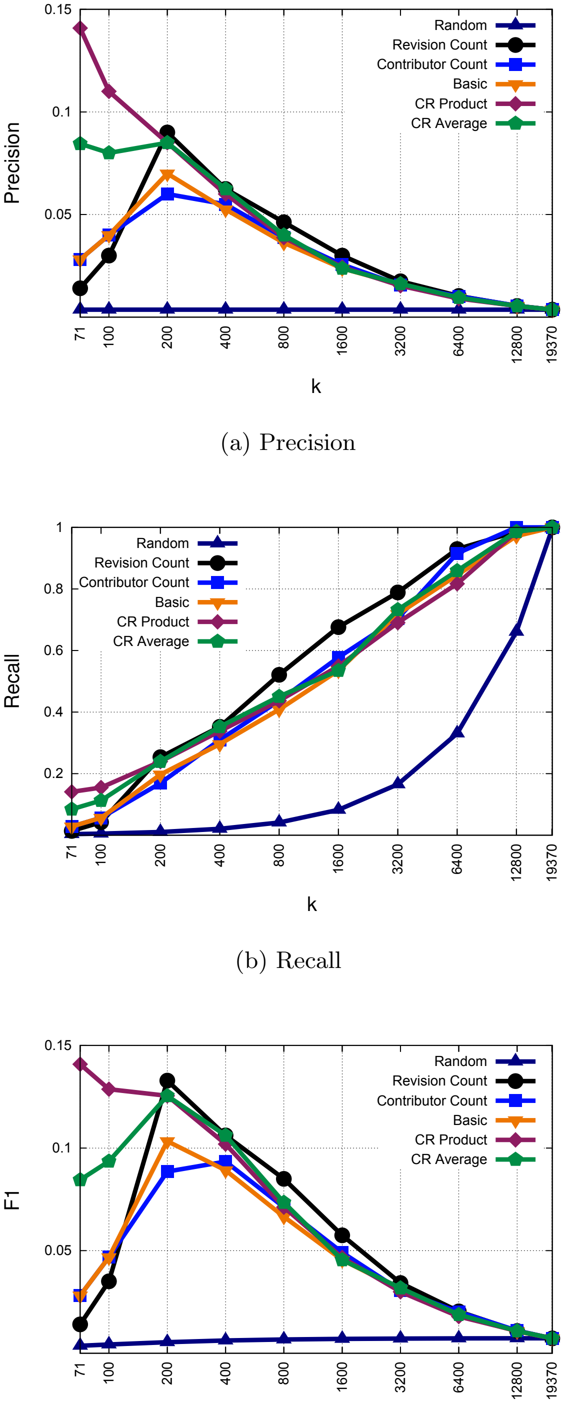
All our models were computed in a relatively short time. The Basic model did not require any extensive computational power and returned results within a few hundred milliseconds. The CR Average and CR Product models were computed by first initializing the same controversial scores (i.e., 1/# contributors) to all contributors. Once the first set of article controversy scores have been computed, they were used to compute the new values for contributor controversy. This process was repeated until the scores reached convergence.

The CR Average and CR Product models took 149 and 43 iterations to converge in approximately 15 minutes and 10 minutes respectively. Convergence was assumed when the delta changes to the controversy values of both articles and contributors were less than the threshold of 10^{-6}.

In this section, we evaluate and compare the performance of five different models, namely *Revision Count*, *Contributor Count*, *Basic*, *CR Average* and *CR Product* models. We also show the performance of the *Random* model for reference.

4.3.1 Precision-Recall-F1

Figures 4 shows the Precision, Recall and F1 values of the different models for different values of k, ranging from 71 to 19,456. In terms of F1, there is no single model that performs best for all k values. CR Product achieves much better F1 results than the other models for $k < 200$. This is important as we want the top ranked articles to be truly the most controversial ones. For $k < 200$, CR Average is the next best performing model. For $k > 200$, there is no clear distinction among CR Product, CR Average and other models. Interestingly, the Revision Count model performs well at $k = 200$. This is due to a substantial number of controversial articles having large number of revision counts. The figures also reveal that Contributor Count and Basic do not perform well in precision and recall. For smaller k's



(a) Precision

(b) Recall

(c) F1

Figure 4: Precision, Recall and F1 @ k performance, k=71 to 19,456 (in log scale)

(say, $k \leq 100$) , Contributor Count and Basic appear to be better than Revision Count. Revision Count however performs better for larger k's.

These above results suggest that the number of revisions or contributors are not reliable indicators of controversy in comparison with other features used in the CR models. By measuring the amount of disputes between contributor pairs and by adopting the mutual reinforcement principle between article and contributor controversies, our CR models appear to perform better than the rest for small k's.

Note that the precision, recall and F1 values are relatively small. This is mainly due to the large collection size. With only 71 non-zero ATC articles, a random ranking approach will yield Precision at $k{=}\frac{71}{19370} = 0.0037$ and Recall at $k{=}\frac{k}{19370}$ respectively. These values are much worse than what our models can achieve. The small precision, recall and F1 values are also partially due to the incomplete set of labeled controversial articles (those with non-zero ATC) which will be illustrated in Section 5.3.

4.3.2 NDCG

Figure 5 shows the NDCG @ k performance of different models. Unlike Precision-Recall @ k, the NDCG @ k performance curves show very clear separations between different models. The two CR models turn out to be the best using this measure. CR Product outperforms all others for all k values. CR Average is the second best performing model. This result confirms that the CR Product is the best model in identifying highly controversial articles.

The figure also shows clearly that Basic is better than Revision Count for small k (<200) but was surpassed by Basic for larger k's. We believe this is somehow due to the nature of the dataset and plan to investigate this further in other experiments. It is also interesting to observe that, using NDCG @ k, the Basic model is clearly inferior to the CR Average and CR Product models.

Since NDCG takes into account different degrees of controversy in articles provided by ATC labels, it is considered a more accurate measure compared to F1. It is therefore reasonable to conclude that the performance ordering of the models from the best to the worst is { CR Product, CR Average, (Basic or Revision Count), Contributor Count}. Henceforth, we shall report only the NDCG results.

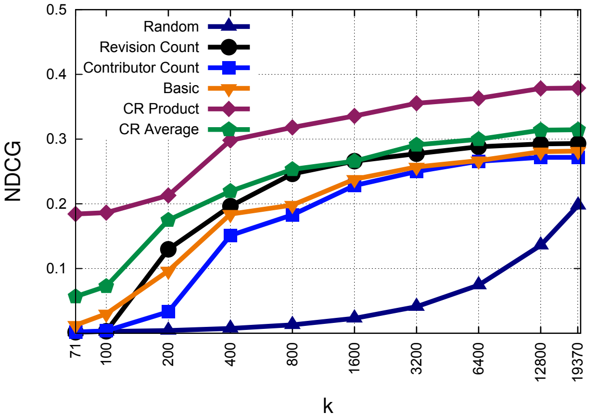


Figure 5: NDCG @ k performance.

4.3.3 Top 20 articles

We now show the top 20 controversial articles returned by the best three models, Basic, CR Average and CR Product in Tables 6, 7 and 8 respectively.

Similar to the Revision Count model (see Table 2) and the Contributor Count model (see Table 3), the Basic model cannot find any labeled controversial articles in the top 20. On the other hand, both the CR Average and CR Product models are able to find 2 and 3 non-zero ATC value articles in the top 20 list, respectively. The highest ATC value article "Holiest sites in Islam" ($ATC = 490$) and two other controversial articles have been found by the CR Product Model. CR Average can only find two articles, namely "Pro-Test" (with $ATC = 58$) "Myron Evans" (with $ATC = 60$).

5. AGE-AWARE MODELS

5.1 Age Functions and Models

In Wikipedia, articles typically undergo several phases. During the construction phase when an article is first created, many edit operations are expected to organize the article content and to work out consensus among the various contributors. Disputes during this phase are rank and file, and do not necessary imply that the article is controversial. The same argument can be applied to a new contributor. Our current Controversy Rank models, however are age-oblivious. As a result, new articles and new contributors may be wrongly assigned high controversy scores.

To overcome the above limitations, we introduce the age-aware versions of CR Product and CR Average, i.e., **age-aware CR Product** and **age-aware CR Average**, by incorporating the age of articles. The main idea behind these age-aware models is to reduce the significance of disputes in new articles while keeping the advantages of the CR models.

We represent the age of an article by the number of revisions that the article has undergone. The age function for articles is given in Equation13.

$$Age(r_k) = \frac{1}{1 + e^{-\frac{1}{2}(rev_k^r - 20)}} \qquad (13)$$

In the above equation, rev_k^r is the number of revisions in article r_k. The value 20 in the exponential term is the median of articles' revision count. The age function returns a value close to zero when rev_k^r is small, 0.5 when rev_k^r is the median value, i.e. $rev_k^r = 20$, and a value close to 1 When rev_k^r is very large.

Using the age function, the revised equations for age-aware Basic are given in Equations 14 and 15.

$$C_k^r = \frac{\sum_{i,j} d_{ijk}}{\sum_i o_{ik}} \times Age(r_k) \qquad (14)$$

$$C_i^u = \frac{\sum_{j,k}(d_{ijk} + d_{jik})}{\sum_{j,k} o_{jk} \times I(i,j,k) + \sum_k o_{ik}} \qquad (15)$$

The controversies of articles and contributors in the age-aware CR models are defined by Equations 16 and 17.

$$C_k^r = \frac{\sum_{i,j} agg[(1 - C_i^u), (1 - C_j^u)] \times d_{ijk}}{\sum_i o_{ik}} \times Age(r_k) \qquad (16)$$

$$C_i^u = \frac{\sum_{j,k}[(1 - C_k^r) \times (d_{ijk} + d_{jik})]}{\sum_{j,k} o_{jk} \times I(i,j,k) + \sum_k o_{ik}} \qquad (17)$$

Note that there is no change to the contributor controversy functions.

5.2 Experiments and Results

5.2.1 NDCG

Figure 6 shows the NDCG @ k for age-aware models and non age-aware models. It shows that the age function actually improves all the models with different degrees. The age-aware CR Product has the most improvement and gives the best performance for all k's.

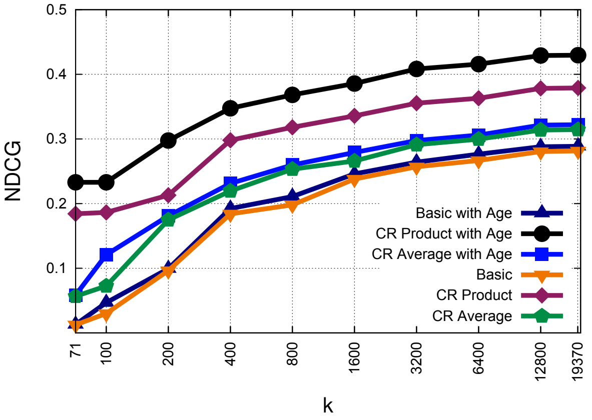


Figure 6: NDCG @ k performance

5.2.2 Top 20 Articles

Tables 9, 10 and 11 show the top 20 controversial articles returned by the age-aware Basic model and CR models. It can be seen that Basic does not improve much from Age function. CR Average and CR Product show improvements compared to their corresponding age-oblivious models. The age aware CR Average is able to rank "Pro-Test" and "Myron Evans" higher than CR Average.

The age-aware CR Product model yields a very impressive performance by locating six non-zero ATC articles in the top 20. It also ranked the top ATC article, "Holiest sites in Islam" at position 8. Most of the non-zero ATC articles listed in top 20 of CR Product model are also very controversial according to their corresponding ATC values.

5.3 Detailed Examples

In this section, we present the detail of some interesting examples that are found by the age-aware CR Product and CR Average models. This illustrates how the two models can more effectively rank controversial articles using the mutual reinforcement principle.

As shown in Tables 10 and 11, the article "Podcast" and "Myron Evans" are the two most controversial articles identified by age-aware CR Average and age-aware CR Product respectively. We therefore examine the two articles as follows.

- *Large portion of disputes:* We observe that these two articles involve a lot of disputes per historical words.

Table 6: Basic Model Top 20 Articles

Rank	Article Name	*ATC*	*ATC* Rank	Rank	Article Name	*ATC*	*ATC* Rank
1	Dominus Illuminatio Mea	0	>71	11	John Howard	0	>71
2	North Marston	0	>71	12	Emma Watson	0	>71
3	Will Young	0	>71	13	Rozen Maiden	0	>71
4	Abingdon School	0	>71	14	Saint Sophia...	0	>71
5	Kamakhya	0	>71	15	Stephen Hawking	0	>71
6	John Dalton	0	>71	16	Queen Elizabeth II Bridge	0	>71
7	Christ Church...	0	>71	17	Aaron	0	>71
8	Podcast	0	>71	18	Kevin Rudd	0	>71
9	Jyotiba	0	>71	19	George Orwell	0	>71
10	Iain Lee	0	>71	20	Our Lady of the...	0	>71

Table 7: CR Average Top 20 Articles

Rank	Article Name	*ATC*	*ATC* Rank	Rank	Article Name	*ATC*	*ATC* Rank
1	John Howard	0	>71	11	Globe Theatre	0	>71
2	Podcast	0	>71	12	Aaron	0	>71
3	Iain Lee	0	>71	13	Anton Chekhov	0	>71
4	Zt”l	0	>71	14	Pro-Test	58	10
5	Stephen Hawking	0	>71	15	Myron Evans	60	9
6	George Orwell	0	>71	16	Our Lady of the...	0	>71
7	Emma Watson	0	>71	17	Robert Hooke	0	>71
8	11-Sep	0	>71	18	St. Dr. Seuss	0	>71
9	Jyotiba	0	>71	19	John Dalton	0	>71
10	Andrew Adonis...	0	>71	20	John Locke	0	>71

Table 8: CR Product Top 20 Articles

Rank	Article Name	*ATC*	*ATC* Rank	Rank	Article Name	*ATC*	*ATC* Rank
1	Zt”l	0	>71	11	Bishop of Salisbury	0	>71
2	Myron Evans	60	9	12	First Baptist Church...	3	52
3	Solomon’s Temple	0	>71	13	Holiest sites in Islam	490	1
4	University College Record	0	>71	14	San Lorenzo...	0	>71
5	Nassau Presbyterian Church	0	>71	15	Guy Davenport	0	>71
6	Shrine of St. Margaret...	0	>71	16	Bonn Minster	0	>71
7	Bishop of Worcester	0	>71	17	Temple Rodef Shalom	0	>71
8	Yell Group	0	>71	18	Ta Som	0	>71
9	St Volodymyr’s Cathedral	0	>71	19	Romanian Orthodox...	0	>71
10	Ashtalakshmi Kovil	0	>71	20	Italo-Greek Orthodox...	0	>71

Table 9: Age-aware Basic Model Top 20 Articles

Rank	Article Name	*ATC*	*ATC* Rank	Rank	Article Name	*ATC*	*ATC* Rank
1	Will Young	0	>71	11	Saint Sophia...	0	>71
2	Abingdon School	0	>71	12	Stephen Hawking	0	>71
3	Kamakhya	0	>71	13	Queen Elizabeth II Bridge	0	>71
4	John Dalton	0	>71	14	Aaron	0	>71
5	Podcast	0	>71	15	Kevin Rudd	0	>71
6	Jyotiba	0	>71	16	George Orwell	0	>71
7	Iain Lee	0	>71	17	Our Lady of the...	0	>71
8	John Howard	0	>71	18	Thomas Hobbes	0	>71
9	Emma Watson	0	>71	19	Globe Theatre	0	>71
10	Rozen Maiden	0	>71	20	Gary Glitter	0	>71

Table 10: Age-aware CR Average Top 20 Articles

Rank	Article Name	*ATC*	*ATC* Rank	Rank	Article Name	*ATC*	*ATC* Rank
1	Podcast	0	>71	11	Andrew Adonis...	0	>71
2	Iain Lee	0	>71	12	Anton Chekhov	0	>71
3	Stephen Hawking	0	>71	13	Our Lady of the...	0	>71
4	George Orwell	0	>71	14	Robert Hooke	0	>71
5	John Howard	0	>71	15	Dr. Seuss	0	>71
6	Emma Watson	0	>71	16	Thomas Hobbes	0	>71
7	Jyotiba	0	>71	17	John Locke	0	>71
8	Globe Theatre	0	>71	18	Pro-Test	58	10
9	Aaron	0	0	>71	John Dalton	0	>71
10	Myron Evans	60	9	20	Oscar Wilde	0	>71

Table 11: Age-aware CR Product Top 20 Articles

Rank	Article Name	*ATC*	*ATC* Rank	Rank	Article Name	*ATC*	*ATC* Rank
1	Myron Evans	60	9	11	Temple Rodef Shalom	0	>71
2	Solomon's Temple	0	>71	12	Romanian Orthodox...	0	>71
3	Bishop of Worcester	0	>71	13	Ashtalakshmi Kovil	0	>71
4	Yell Group	0	>71	14	Italo-Greek Orthodox...	0	>71
5	St Volodymyr's Cathedral	0	>71	15	City Harvest Church	27	18
6	Bishop of Salisbury	0	>71	16	Macedonian Orthodox...	61	8
7	First Baptist Church...	0	>71	17	Oxford University Conservative...	1	64
8	Holiest sites in Islam	490	1	18	Church of Kish	21	24
9	Guy Davenport	0	>71	19	Iain Lee	0	>71
10	Bonn Minster	0	>71	20	Waldegrave School...	0	>71

Among 381,567 historical words in "Podcast", 380,142 (99.6%) are disputed words. Similarly, among 6431 historical words in "Myron Evans", 6191 (96.3%) are disputed. The proportions of disputed words for these two articles are significantly larger than 69.5%, the average proportion of disputed words for the articles in the dataset.

- *More disputes between less controversial contributors:* For "Podcast", a significant amount of disputes occurred between the two pairs of users: (a) user 210.213.171.25 and user Jamadagni; and (b) user 68.121.146.76 and user Yamamoto Ichiro. 11.15% and 8.12% of the disputed words in the article came from pairs (a) and (b) respectively. Users 210.213.171.25, Jamadagni, 68.121.146.76 and Yamamoto Ichiro are ranked 100244th, 10245th, 100242nd and 81346th respectively. The relatively low controversial ranks for the two users with IP addresses suggest that they are the less controversial contributors according to Age-aware CR Average. However, the inverse controversy scores of these users lead to the higher controversial "Podcast" article. For "Myron Evans", the top user pairs in dispute are; (c) user Solmil and user Rich Farmbrough; and (d) user Mwkdavidson and user Mathsci. 13.45% and 11.35% of the disputed words in the article came from pairs (c) and (d) respectively. Somil, Rich Farmbrough, Mwkdavidson and Mathsci are ranked 91268th, 143464th, 86412nd and 99883rd respectively. These are the relatively low controversial users, hence causing the high controversial score for the article.

- *More revisions:* Both articles have a reasonable number of revisions. "Podcast" has a huge number of revisions, i.e., 5344, and "Myron Evans" has 277 revisions. In other words, they are not considered new articles by the age-aware models.

These two top controversial articles are ranked high by not only a single model but also our other proposed models. Age-aware Basic model ranks "Podcast" and "Myron Evans" 5th and 134th respectively. Age-aware CR Product ranks "Podcast" 389th and age-aware CR Average ranks "Myron Evans" 10th. These rank positions are relatively small compared with the total number of articles ranked (19,456). This shows that all our proposed models achieve quite similar results for these two articles.

It is also interesting to point out that "Podcast" has zero *ATC* values, meaning that it has never been tagged as controversial before. However when reading the edit history and talk page of this article, we can see a lot of vandalism among its contributors. Since the article covers podcasting, a modern form of entertainment, it tends to attract anonymous user contribution, as well as anonymous vandalism. The word "vandalism" appears 80 times in the edit comments.

A contributor regarded as very controversial by our age-aware CR Product model is "Ak8243". He is ranked 3rd by age-aware CR Product, 12th by age-aware CR Average, and 120,846th by Basic. Looking at his talk page, one can see that he has instigated a lot of disputes with other contributors and his talk page is full of complaints from others. For example, his created article "Burger Travel Service" is considered by others as a blatant advertisement with no reliable source. Some users consider this article a spam. "Ak8243" has also caused conflicts with other users by trying to remove the notice "Articles for deletion " that others have added to his articles. As a result,"Ak8243" is regarded as a vandal by some contributors and his article "Burger Travel Service" was finally removed from Wikipedia.

6. RELATED WORK

Finding controversial Wikipedia articles is a very new but challenging research problem. In our literature review, we could not find much work done in this area. Much of the existing research on Wikipedia data focuses on its article reputation and quality, which are two closely related concepts [1, 3, 7, 10, 11, 17, 18],

Lih [10] is among the first who analysed the creation and evolution of Wikipedia articles. He postulated that rigor (total number of edits) and diversity (total number of unique authors) are important features of reputable articles. The median values of these two features were estimated experimentally and then used to identify reputable articles. He also showed that citations from other established media has driven public attention directly to certain articles of Wikipedia, and has improved their reputation subsequently. Anthony et al. observed that registered users who contribute frequently and non-registered users who contribute infrequently are the two groups of users who produce articles with good quality in Wikipedia[3].

In our early work [7, 11], we proposed several models to measure Wikipedia article quality and contributor authority. These models do not interpret the article content. Instead, they rely on the edit histories of articles that record the collaborations among contributors. The models were designed based on the mutual reinforcement principle: "good contributors usually contribute good articles and good articles are contributed by good authors". Depending on the type of contribution (e.g. authorship, reviewership), we derived the different models. Our experiments showed that the model based on reviewership gives the best article quality prediction. In [1], Adler and Alfaro assigned each contributor a cumulative reputation according to text survival and edit survival of their revisions. The contributor would gain reputation if his/her edit is long-lived. In contrast, short-lived edits would gain less or even negative reputation. The experiments showed that content contributed by low-reputation authors are more likely to suffer from poor quality and removals.

The above approaches to assess Wikipedia article quality and reputation, however, cannot be directly applied to measuring controversy of articles because quality (as well as reputation) and controversy are two distinct concepts. The former concerns with how well the article is written while the latter is related to disputes due to the article topic and/or contributors editing the article.

The work closest to finding controversial articles is done by Kittur [9]. Kittur et al. used a set of page metrics (including number of revisions, page length, number of unique editors, links from other articles, etc.) as the features of Wikipedia articles to train a Support Vector Machine (SVM) classifier for assigning a score to each controversial article [13]. While their experiments found that the learnt classifier was able to rank the controversial articles consistent with their actual degrees of controversy, the method was evaluated using a small set (272 articles) of Wikipedia articles including controversial articles only. In contrast, we have adopted a non-supervised approach to rank controversial articles. We have also chosen to link controversial articles with controversial contributors. Our experiments also involve a much larger article set including articles with and without dispute tags.

7. CONCLUSIONS

Finding controversial articles in Wikipedia is an interesting and challenging problem. A good solution to this problem can greatly improve the ways controversial topics can be automatically detected, Wikipedia articles can be ranked in search results, etc..

In this paper, we have presented several novel models to rank Wikipedia articles and contributors by their degrees of controversy. The three proposed models, Basic, CR Average and CR Product, share the common idea of measuring disputes among contributors in their edited articles. The CR models are extensions of the Basic model by considering the mutual relationship between controversy of articles and that of contributors. Our first set of experiments results show that our models work well and outperform the baselines for the top ranked articles. We further improve our models by considering the age of articles and contributors. The performance gain by the age-aware models is verified by the second set of experiments.

As part of our future research, we shall look into choosing the parameters used in our models so as to optimize the ranking results. It is also important to study how the models can be extended to rank the entire Wikipedia collection efficiently. We are also interested to study the theoretical properties such as proof of convergence, relative performance with large/small numbers of articles, etc., of our proposed models. Finally, we plan to conduct a comprehensive user evaluation on the discovered controversial articles so as to compare the models from the user perspective.

8. REFERENCES

[1] B. Adler and L. de Alfaro. A content-driven reputation system for the Wikipedia. In *Proc. of WWW'07*, pages 261–270, 2007.

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

[3] D. Anthony, S. Smith, and T. Williamson. Explaining quality in internet collective goods: Zealots and good samaritans in the case of Wikipedia, 2005. Hanover : Dartmouth College.

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

[5] S. Flexner and L. Hauck. *Random House Unabridged Dictionary*. Random House, New York, NY, 2nd edition, 1993.

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

[7] M. Hu, E.-P. Lim, A. Sun, H. Lauw, and B.-Q. Vuong. Measuring article quality in wikipedia: Models and evaluation. In *Proc. of CIKM'07*, 2007.

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

[9] A. Kittur, B. Suh, B. Pendleton, and E. Chi. He says, she says: conflict and coordination in wikipedia. In *Proc. of SIGCHI'07*, pages 453–462, 2007.

[10] 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.

[11] E.-P. Lim, B.-Q. Vuong, H. Lauw, and A. Sun. Measuring qualities of articles contributed by online communities. In *Proc. of WI'06*, December 2006.

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

[13] A. Smola and B. Schölkopf. A tutorial on support vector regression. statistics and computing. *Statistics and Computing*, 14:199–222, 2004.

[14] J. Voss. Measuring wikipedia. In *Proc. of International Confererence of the International Society for Scientometrics and Informetrics*, Stockholm, Sweden, July 2005.

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

[16] Wikipedia. Wikipedia, 2007. `http://en.wikipedia.org/wiki/Wikipedia` Accessed on April 2007.

[17] Wikipedia. Wikipedia in academic studies, 2007. `http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_in_academic_studies` Accessed on April 2007.

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