Automatically Assessing Wikipedia Article Quality by Exploiting Article–Editor Networks

Xinyi Li† ‡, Jintao Tang†, Ting Wang†, Zhunchen Luo§ and Maarten de Rijke‡

† National University of Defense Technology, Changsha, China
tangjintao, tingwang@nudt.edu.cn
‡ University of Amsterdam, Amsterdam, The Netherlands
x.li, derijke@uva.nl
§ China Defense Science and Technology Information Center, Beijing, China
zhunchenluo@gmail.com

Abstract. We consider the problem of automatically assessing Wikipedia article quality. We develop several models to rank articles by using the editing relations between articles and editors. First, we create a basic model by modeling the article-editor network. Then we design measures of an editor's contribution and build weighted models that improve the ranking performance. Finally, we use a combination of featured article information and the weighted models to obtain the best performance. We find that using manual evaluation to assist automatic evaluation is a viable solution for the article quality assessment task on Wikipedia.

1 Introduction

Wikipedia is the largest online encyclopedia built by crowdsourcing, on which everyone is able to create and edit the contents. Its articles vary in quality and only a minority of them are manually evaluated high quality articles.[1] Since manually labeling articles is inefficient, it is essential to automatically assess article quality. Content quality criteria are known to help retrieval; in a web setting they are often based on link structure [7, 8] but in the setting of social media and collaboratively created content, content-based features are often used [11]. Here, we study the quality assessment of Wikipedia articles by exploiting the article-editor network. We view this task as a ranking problem. Our task is motivated by the assumption that automatic procedures for assessing Wikipedia article quality can help information retrieval that utilizes Wikipedia resources [2] and information extraction on Wikipedia [13] to obtain high quality information.

There have been different approaches to the content quality assessment problem. One branch of research uses simple metrics, such as article length, number of links and citations etc. [1, 6, 9, 12]. These authors do not consider the interactions between editors and articles, which differentiates Wikipedia from traditional encyclopedias. Other work takes into account the network of articles and editors. Hu et al. [4] proposes what they call a probabilistic review model to rank articles. The model is tested on a dataset of only 242 articles. Suzuki and Yoshikawa [10] uses a combination of survival ratio

[1] Only 0.1% of all Wikipedia articles are featured articles.

method and link analysis to score articles. They use relative evaluation metrics to measure the performance of models. It remains to be seen to which degree they can achieve satisfactory ranking results in more realistic settings.

We examine the editing actions of editors and find that the majority of them are field-specific, i.e., they specialize in a certain category of articles. These field-specific editors outnumber all-around editors to a great extent. Since the editor-article networks of different categories only share very few nodes, ranking articles should be done in separate categories. As featured articles are manually-tagged high quality articles, we select them as the ground truth for our task. We develop several models to rank articles by quality. Our first motivation is to see if the importance of a node in the network can indicate quality. So we develop a basic PageRank-based model. Additionally, instead of treating links as equal in the basic model, we tweak the model by putting weights on the links to reflect the difference of editor contributions. Finally, we utilize existing manual evaluation results to improve automatic evaluation. So we incorporate manual evaluation results into our model. We use articles of different quality levels to measure the levels of editors, and then assist ranking.

The experiments carried out on multiple datasets covering different fields show that ranking performance is related to the number of high quality articles we utilize. In particular, the higher the percentage of high quality articles used, the better the ranking performance. We also find that the basic model does not yield satisfactory ranking results, but that using weights boosts performance.

2 Models

We introduce the models and explain how each model is computed, including a baseline model, weighted models, and weighted models with probabilistic initial value.

2.1 Baseline model First, we develop a basic quality model based on Pagerank. PageRank is widely applied for ranking web pages, where pages are seen as nodes and hyperlinks as edges [7]. The node value represents its importance in the network. In our basic model we treat both articles and editors as nodes connected by edges that represent editing relations. For instance, if article A is edited by B then there is a bidirectional edge that connects A and B. The value of the nodes are distributed through the edges during each iteration of the PageRank computation. As shown in (1), the value of node v is determined by nodes in the set $U(v)$ that connect to it, where $N(u)$ is the number of edges that point out of node u.

$$PR(v) = (1-d) + d \sum_{u \in U(v)} \frac{PR(u)}{N(u)}. \tag{1}$$

In this basic model, we give all nodes the same initial value and iteratively compute the node value until they converge. The articles will then be ranked by node value.

2.2 Weighted models The baseline model treats edges as equal. However, consider an article that has multiple editors, which is quite common. When the value of the article node is distributed toward its editors during computation, editors that make a higher contribution should get more. There should be a weight to address this difference.

It is therefore necessary to measure how users contribute to article quality and how articles contribute to user authority in return. While it is hard to precisely quantify the contribution, we can use editing actions during an article's history as an approximation. An intuitive measure is to use the edit counts between article and editor as a measure, defined in (2):

$$Contribution1 = \#edits. \tag{2}$$

We define the weighted model based on this equation as the *simple weighted* (SW) model. By further parsing the editing actions, we can obtain a more complex measure that takes different editing behaviors into account, which is defined in (3):

$$Contribution2 = \#insertions + \#deletions + \#replacements. \tag{3}$$

An editor's contribution to an article is the sum of words affected by their editing actions. The editing actions are insertion (insert new content), deletions (delete content) and replacements (insert new content right after deletion), which are shown to have a strong correlation with article quality[5]. As Wikipedia only provides history versions of articles, we obtain the editing actions by comparing adjacent article revisions with a diff-algorithm [3]. We define this model as the *complex weighted* (CW) model. After defining the contribution, we put the contribution value on each edge as the weight. The value of nodes is defined in (4).

$$PR(v) = (1-d) + d\sum_{u \in U(v)} PR(u)\frac{C_{uv}}{\sum C_u}. \tag{4}$$

In this equation the value of node u will be multiplied by the proportion of the weight value C_{uv} against the weight sum $\sum C_u$.

2.3 Weighted models with probabilistic initial value To further improve ranking, we incorporate manual evaluation results into our weighted models. Our hypothesis is that featured articles and other articles have different levels of editors. Using articles of different quality to differentiate editors' levels may improve article ranking.

To do so, we simply give articles different initial values before computation. Their value will then be distributed to editors through editing relations. An article's initial value is determined by its probability of being high quality. We assign an initial value of 1.0 to featured articles because they have a probability of 100% to be high quality articles. Likewise, we set the initial value of other articles as the proportion of featured articles to all articles in that particular category. We set the initial value of editors as 0. We define the models as the *simple weighted probabilistic* (SWP) model and the *complex weighted probabilistic* (CWP) model based on different contribution measures.

3 Experimental Setup

3.1 Datasets We select three categories from an English Wikipedia dump[2] as a case study. These categories cover different fields and contain both high quality articles and articles of unknown quality. The statistical information of the articles in these categories

[2] Data dump of March 15, 2013, fetched from https://dumps.wikimedia.org/.

Table 1. Statistics of datasets.

Category	#articles	#editors	#featured articles
Chemistry	7,796	392,055	36
Meteorology	4,218	187,637	138
Geography	38,543	1,360,508	180

is shown in Table 1. We find that most editors specialize in one field, and only a minority of them are all-around editors. Therefore the article-editor networks of different categories only share a tiny proportion of common nodes. Based on this structure of the article-editor network, we will apply ranking by category.

3.2 Metrics We assess article quality by ranking. Since featured articles are the best quality articles on Wikipedia, they are frequently used as the gold standard to measure ranking performance. However, common metrics such as RMSE are not suitable for this task as Wikipedia does not give a specific ranking for featured articles. We consider recall scores at the first N items in the result set, as well as precision-recall curves.

3.3 Parameter settings In the baseline model and weighted models, we initially assign 1.0 to all nodes and iteratively compute their values. The iterations can be halted for any desired mean error of the ranking being less than 0.01. For the SWP and CWP models, we will initialize them using probabilistic values as explained earlier.

4 Experimental Evaluation

We address two main research questions. We contrast our four methods, i.e., the Baseline method, the simple weighted model (SW), the complex weighted (CW) model, as well as two variants with probabilistic initial values (SWP, CWP). But first we examine the impact of the number of featured articles used for initialization in SWP and CWP. We want to find out how this number affects ranking performance.

Table 2 shows that in most cases, the more featured articles used for initialization in SWP or CWP, the better the ranking performance. We notice a few exceptions to this finding, especially in categories that have more featured articles. This is because many of the featured articles used in initialization are ranked atop, reducing the chance for other articles in the ground truth to rank high. Still, by using all featured articles for initialization we achieve the best recall performance.

Next, we compare SWP and CWP in this best case with the previous models in Figure 1. To determine whether the observed differences between two models are statistically significant, we use Student's t-test, and look for significant improvements (two-tailed) at a significance level of 0.99. We find that both SWP and CWP statistically significantly outperform other models in all categories. We also note that the SWP model performs better than the CWP model in most cases, which is contrary to the previous experiments where the complex contribution measure yields better results. The best ranking performance is achieved by the SWP model when using all available high quality articles in initialization. And the recall levels are up to an applicable value.

Table 2. Recall (N) of SWP and CWP in different categories.

featured%	r@100		r@200		r@300		r@400	
	SWP	CWP	SWP	CWP	SWP	CWP	SWP	CWP
chemistry								
25%	.556	.363	.767	.667	.867	.793	.440	.874
50%	.644	.378	.778	.694	.861	.833	.972	.883
75%	.756	.400	.911	.744	.956	.911	1.000	.944
meteorology								
25%	.111	.092	.246	.175	.365	.317	.498	.421
50%	.101	.103	.274	.165	.438	.346	.607	.486
75%	.140	.114	.346	.200	.517	.357	.703	.514
geography								
25%	.173	.086	.342	.168	.426	.283	.496	.369
50%	.163	.069	.357	.182	.497	.317	.562	.422
75%	.149	.051	.376	.162	.518	.327	.596	.407

E.g., the recall value at $N = 200$ is 0.756 in geography, meaning that the 180 featured articles in that category have a probability of 75.6% to appear in the top-200 list.

We also notice that ranking performance is related to the number of featured articles in each category. E.g., chemistry, which has the fewest featured articles, is higher in precision than other categories at a given recall level. Meanwhile, the curves of meteorology and geography both experience a rise and then gradually descend. This is because at the top of the result list are mostly featured articles, and then false positives are appearing at an increasing speed in the list, causing the curve to go downwards.

Our SWP model has achieved the best precision/recall performance and is far better than using content-based features. For instance applying our SWP model in chemistry category gives a recall score of 0.889 out of the top 100 items, while using content-based features in Blumenstock [1] only yields 0.306. Our evaluation metrics are also more applicable for ranking purpose than the relative measures used in Suzuki and Yoshikawa [10], so that we can apply our model in a practical setting.

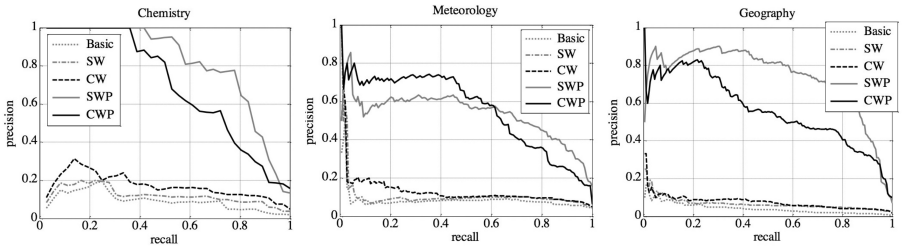

Fig. 1. Precision-recall curves for the baseline (Basic), simple weighted (SW), complex weighted (CW), simple weighted probabilistic (SWP), complex weighted probabilistic (CWP) model.

5 Conclusion

We have developed several models for estimating Wikipedia article quality based on the article-editor network. They include a basic model, a weighted model, which addresses the difference of editors' contributions, and probabilistic weighted models incorporating manual evaluation results. The experimental results show that by using featured articles, we are able to differentiate editor levels and then improve ranking performance.

Additionally, the baseline model we considered (based on PageRank) does not yield satisfactory ranking results, but when we put weights on the links, the ranking results receive a boost. The improvements are not as significant as using featured articles.

Summarizing, the combination of existing manual evaluation results (featured articles) with the article-editor network yields a state-of-the-art solution for assessing article quality. For future work, we will improve our model by adding features of editors, and also conduct a systematic comparison with the methods presented in [4, 10].

Acknowledgments. This research was partially supported by the National Natural Science Foundation of China (Grant No. 61472436 and 61202337), the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 312827 (VOX-Pol), the Netherlands Organisation for Scientific Research (NWO) under project nrs 727.011.005, 612.001.116, HOR-11-10, 640.006.013, 612.066.930, the Center for Creation, Content and Technology (CCCT), the Dutch national program COMMIT, the ESF Research Network Program ELIAS, the Elite Network Shifts project funded by the Royal Dutch Academy of Sciences (KNAW), the Netherlands eScience Center under project number 027.012.105, the Yahoo! Faculty Research and Engagement Program, the Microsoft Research PhD program, and the HPC Fund.

6 References

[1] J. E. Blumenstock. Size matters: word count as a measure of quality on Wikipedia. In *WWW*, 2008.

[2] T. Cassidy, H. Ji, L.-A. Ratinov, A. Zubiaga, and H. Huang. Analysis and enhancement of wikification for microblogs with context expansion. In *COLING*, 2012.

[3] O. Ferschke, T. Zesch, and I. Gurevych. Wikipedia revision toolkit: efficiently accessing Wikipedia's edit history. In *ACL*, 2011.

[4] M. Hu, E.-P. Lim, A. Sun, H. W. Lauw, and B.-Q. Vuong. Measuring article quality in Wikipedia: models and evaluation. In *CIKM*, 2007.

[5] X. Li, Z. Luo, K. Pang, and T. Wang. A lifecycle analysis of the revision behavior of featured articles on Wikipedia. In *ISCC*, 2013.

[6] A. Lih. Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. In *ISOJ*, 2004.

[7] B. Liu. *Web Data Mining*. Springer, Heidelberg, 2007.

[8] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford University, 1999.

[9] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Assessing information quality of a community-based encyclopedia. In *IQ*, 2005.

[10] Y. Suzuki and M. Yoshikawa. Assessing quality score of Wikipedia article using mutual evaluation of editors and texts. In *CIKM*, 2013.

[11] W. Weerkamp and M. de Rijke. Credibility-based reranking for blog post retrieval. *Information Retrieval Journal*, 15(3–4):243–277, June 2012.

[12] D. Wilkinson and B. Huberman. Cooperation and quality in Wikipedia. In *WikiSym*, 2007.

[13] F. Wu and D. S. Weld. Open information extraction using Wikipedia. In *ACL*, 2010.