Clickcount-Weighted PageRank Algorithm for Finding High Quality Documents

He LIU[†] and Mizuho IWAIHARA[‡]

† 早稲田大学大学院情報生産システム研究科

‡ 808-0135 福岡県北九州市若松区ひびきの2-7-N251

E-mail: † liuhe@akane.waseda.jp, ‡ iwaihara@waseda.jp

Abstract Finding high quality documents has been discussed, since users always want to not just relevant documents, but those of high quality and informative. Incoming links of a web document can be viewed as the quality assessment by document authors. On the other hand, detailed click count information on each page or link can be viewed as popularity and informativeness judgement by site viewers (users).However, existing link analysis algorithms, including HITS and PageRank, exploiting static link connectivity between pages, which is not very adequate. In this paper, we propose an algorithm called CWPR (Clickcount-Weighted PageRank), which extends the original PageRank algorithm by smoothly integrating link structure and click count distribution, to score quality of documents. Our evaluations on finding featured articles and good articles of English Wikipedia show that our WPR algorithm shows better performance on a large Wikipedia corpus than algorithms that utilize link graphs or click counts only, and clickcount-weighted HITS algorithms.

Keyword Link analysis, Document assessment, Click graph, PageRank

1. Introduction

Since information retrieval systems came into being, ranking of retrieval results has always been a question for researchers to solve. In modern information retrieval systems, the result of one web query can contain thousand millions of entries of pages, while in most cases, users can only read through or even looking through very a few of them. Therefore, the ranking of retrieval results is quite significant. With a good ranking, users can find articles of high quality and informativeness with less effort.

Many methods based on link analysis on web pages have been developed for evaluating the significance of hyperlink structure [8, 13]. The HITS algorithm computes hub and authority scores for each page through mutual recursion [6]. The PageRank algorithm is based on the random surfer model and a PageRank score on a page represents likelihood that a user is on the page [7].However, only static link connectivity between pages is not very adequate to acquire a very solid evaluation on the quality of documents. Thus, in this paper, we discuss smoothly integrating link structure and click count distribution to evaluate the quality of documents.

To achieve this goal, we extend the original PageRank model by integrating link structure and click count distribution, and we proposed CWPR (Clickcount-Weighted PageRank). Although there are various versions of PageRank, such as Topic-sensitive PageRank, Weighted PageRank [9, 15], they do not utilize counts on links, which can be viewed as the votes from users to a page. On the other hand, the link itself can be viewed as votes on pages by users who visited the same page. By integrating the quality evaluations of pages through both link graph and click graph, the biases caused by document authors can be reduced. So our method is expected to exploit the benefits of both link structure-based and click-based approaches. The click count-weighted HITS algorithm [16] is the first attempt to extend the HITS algorithm by integrating click counts and link structure. This paper explores extension of PageRank by these settings.

For experimental evaluations, we choose English Wikipedia articles to observe the result of our algorithm [5]. The English version Wikipedia holds over one million articles of different quality levels, where the quality control is essential. In Wikipedia, high-quality articles assessed by Wikipedia editors will be selected as featured articles by certain criteria [10]. Thus such featured articles are chosen to be the gold standard for our quality evaluation of pages. Also, we use original PageRank and the click count-weighted HITS algorithm [16] to compare with our proposed method.

For click count data, we utilize publicly available transition data from Wikipedia of May 2018[14]. This monthly-generated data is extracted from the server log for the English desktop version of Wikipedia, which contains aggregated and anonymized page requests in the form of (referrer, resource)-pairs. From the click data and the dump xml file of English Wikipedia, we create directed graphs - Wikipedia click graphs where nodes are articles, edges represent links between articles and click count of a link will be partly influence the link's edge weight. Our evaluation on finding articles of high quality in English Wikipedia shows that our CWPR shows better performance on the large corpus than the algorithms that utilized link graphs or click graphs only and the click count-weighted HITS algorithm.

The rest of this paper is organized as below: In Section 2, we introduce related work about ranking retrieval result. In Section 3, we introduce our proposed method in details. In Section 4, we introduce our experiments on English Wikipedia datasets. In Section 5, we make a conclusion and discuss our future work.

2. Related work

Many methods base on link analysis have been proposed. The original HITS algorithm computes hub and authority scores for each page through mutual recursion [6], while the original PageRank algorithm computes PageRank score for each page according to the incoming link count of the page iteratively.

There is also another approach of document quality evaluation: by informativeness and comprehensiveness. Through numerical features such as term distribution, document length or relationships with other documents, the quality of one document can be evaluated.

Blumenstock showed how word count can be utilized as a proxy for article quality [2]. Calzadatried to evaluate the quality of Wikipedia articles by authors and editing history of articles [3]. Suzuki and Yoshikawa proposed a method to identify Wikipedia articles of good quality by mutually evaluating editors and texts [12]. De La Robertie proposed to use the collaboration network of Wikipedia articles to measure article quality [11].

However, these features may not lead to the same result as readers' perceptions. Here, the readers' feedbacks are important indicators of document quality, which are available in various forms, such as click counts or page views, likes or comments and retweets [1]. In the case of Wikipedia, since there exists publicly-available data such as edit histories of articles, anonymized click counts, and quality evaluation of articles is accessible through Wikipedia grading scheme.

3. Proposed algorithm

3.1 Original PageRank algorithm

The original PageRank algorithm is based on a very simple idea defined by a very concise formula. The idea is such that a page is important if it is pointed to by other important pages [7]. The definition is:

$$\mathrm{PR}(i) = \sum_{j \in B_i} \frac{\mathrm{PR}(j)}{\mathrm{L}(j)} \tag{1}$$

Here, $\mathrm{PR}(i)$ is the PR value (or PageRank score) of node i. B_i is the set of nodes having links pointing to node i. $L(j)$ is the number of the outgoing links from node j.

To deal with the loop problem and other problems like dangling nodes, a random teleportation is added (the random surfer will not only follow the link structure but also may go to some node directly and randomly). Correspondingly, the PageRank score is modified to the following:

$$\mathrm{PR}(i) = \frac{1-d}{N} + d \sum_{j \in B_i} \frac{\mathrm{PR}(j)}{L(j)} \tag{2}$$

Here, N is the number of nodes in the graph, and d is the damping factor. In this paper, d is set as the recommended value 0.85. This is a normalized version, meaning that the PR value is the probability of finding the random surfer on the node and the summation of the PR value of all nodes are always 1.

The PageRank algorithm is evaluated by the following:

$$\mathrm{PR}_{k+1}(i) = \frac{1-d}{N} + d \sum_{j \in B_i} \frac{\mathrm{PR}_k(j)}{L(j)} \tag{3}$$

The PR value of node i in the $(k+1)$-th iteration is calculated by the PR values of nodes in B_i in the k-th iteration. The overall process is as follows:

1) Equally distribute the same PR value to all nodes;
2) Use (3) to calculate the PR value for each node;
3) Normalize the PR values of all nodes;
4) Check if the whole result has reached the finish condition. If yes, to 2) again. Else, to 5);
5) Rank the nodes by PR values;

Here, the finish condition depends on the task, either max iteration limit or tolerance of difference between two iteration is popularly used.

As the PR value becomes larger, the corresponding node becomes more important.

3.2 Click count-weighted HITS

We refer [16] for the definition of click count-weighted HITS algorithm. In contrast to the original HITS algorithm and other link analysis-based algorithms, this modified HITS algorithm uses both link and click counts for hub and authority scores. So we compare this algorithm with our proposed method. According to [16], the Click count-weighted HITS algorithms are defined as following:

Version1:

$$auth[i] = \sum_{j=1}^{n} log(hub[j] * w[j][i] + 1)$$

$$hub[i] = \sum_{j=1}^{n} log(auth[j] * w[i][j] + 1)$$

(4)

Here, some notations are introduced:

$auth[i]$: the auth score of article i.

hub[i]: the hub score of article i.

e_{ij}: the number of links from article i to article j.

n: the number of nodes.

w[i][j]: the click counts of a link from article i to article j, will be 0 if there is no link from i to j.

Version2:

$$auth[i] = \sum_{j=1}^{n} (hub[j] * e[j][i] + k * p * log(w[j][i] + 1))$$

$$hub[i] = \sum_{j=1}^{n} (auth[j] * e[i][j] + k * p * log(w[i][j] + 1))$$

(5)

Here, two new parameters k and p are introduced, such that k is a coefficient determining how much the log-damped click count affects the final score and p is defined as following:

$$p = \frac{\sqrt{\frac{1}{n}}}{AVG_{ij}(\log(w[i][j]+1))} \tag{6}$$

3.3Clickcount-Weighted PageRank(CWPR)

The original PageRank is based on link graph only, which is not ideal in a situation where click count distribution is available. Thus, we extend the original PageRank by smoothly integrating link counts and click counts.

Before introducing the proposed method, we first need to introduce several notations and definitions.

$L = [e_{ij}]$: Link adjacency matrix. The element $e_{ij} \geq 0$ equals the number of links from article i to article j.

N: The number of articles in the dataset.

$C = [c_{ij}]$: Click count matrix. Here, $c_{ij} \geq 0$ is the count of clicks on the link from article i to article j.

A_j : The index of articles which have an incoming link from article j.

B_j: The index of articles which have an outgoing link to article j.

d: $0 < d < 1$ is a dumping factor.

With notations above, we now introduce our proposed method. Since the PageRank algorithm is based on the random surfer model [7], we can understand it from the point of view of probability. In the normalized version of PageRank, the PR value of a node is the probability of finding the random surfer on the node. So the PageRank score of article i is:

$$\mathrm{PR}(i) = (1-d)W(i) + d\sum_{j \in B_i} P(i|j)\,\mathrm{PR}(j) \tag{7}$$

Here the PR value of node j is distributed by $\mathrm{P}(i|j)$, the conditional probability of any user on j going to i, which is estimated as:

$$P(i|j) = \frac{f_{ji}}{\sum_{k \in A_j} f_{jk}} \tag{8}$$

Since the click count varies within a wide range in the graph, we need to log-damp the click counts. We combine the link counts and log-damped click counts with the balancing factor $0 \leq \gamma \leq 1$, as follows:

$$f_{ji} = (1-\gamma)\, e_{ji} + \gamma\, log(c_{ji} + 1) \tag{9}$$

For the teleportation probability, it is distributed by node weight $W(i)$, which is defined as follows:

$$W(i) = \frac{c_{ex-i}}{c_{ex}} * 0.5 + \frac{0.5}{N} \tag{10}$$

Here, the c_{ex-i} is the click counts of external links to node i, c_{ex} is the total click counts of external links to the graph.

4. Experiments

4.1 Datasets

4.1.1 Click count information

For the click count information, we choose the tsv file of the clickstream file generated in May 2018[17].This monthly-generated data file is extracted from the server log for the English desktop version of Wikipedia, which contains aggregated and anonymized page requests in the form of (referrer, resource)-pairs. Also, the times of occurrence of pairs are recorded, while records caused by bots and web crawlers, as well as transitions occurring less than 10 times, are removed. There are three types of (referrer, resource)-pairs as follows:

1. Link: if the referrer and resource are both articles and the referrer article have a link to the resource.
2. External: if the referrer host does not match the URI pattern en(.m)?.wikipedia.org.
3. Other: if the referrer and resource are both articles but the referrer does not link to the resource. This can happen when clients search or spoof their

refer.

Here, we only utilize (referrer, resource)-pairs whose type is “Link” or “External”, which means that only click counts of links within English Wikipedia and click counts of links from external sites are taken into consideration.

The counts of links within English Wikipedia are utilized to determine the weights of edges while the counts of external links are utilized to determine the weights of nodes. The total click graph extracted from this data file is $G_{whole-c}$.

4.1.2 Article sets

We extracted 100 sub-graphs of 10 groups from the graph $G_{whole-c}$ of different sizes. For each group i, i=0,…,9, we set a minimum node count min_i and a maximum node count max_i, such that $max_i = min_i$+50000 and min_i =100000*i. For each group, 10 sub-graphs of the node count range [min_i, max_i] are generated, by the following process:

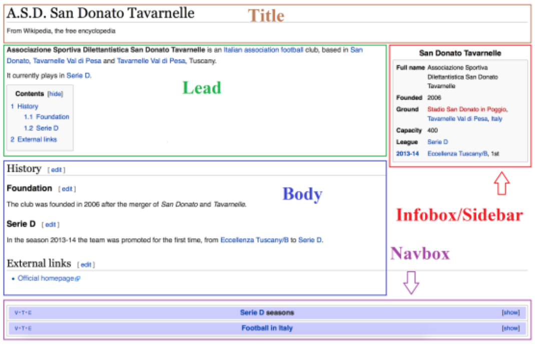
1) Choose a root node by random.
2) By breadth-first-search, nodes adjacent to the current node set are added, until the node count becomes between min_i and max_i, or no more nodes can be added.
3) If there are not enough nodes, go back to Step 1.

The set of the sub-graphs of group i is represented as S_i. The largest sub-graph used in our experiments contains about one million nodes.

4.1.3 Link information

Since the click count file only contains links counted at least once, links not clicked during the collection period are not included. Also, there exist parallel links between two articles. To correctly count links between articles into e_{ij}, we use the widely used tool Wiki-extractor, which can preserve the link tags in the “Lead” and “Body” parts of a Wikipedia page. Thus, the link counts of our dataset reflect all the links in the “Lead” and “Body” parts.

Here, we use the page structure from [4].



A.S.D. San Donato Tavarnelle

From Wikipedia, the free encyclopedia

Associazione Sportiva Dilettantistica San Donato Tavarnelle is an Italian association football club, based in San Donato, Tavarnelle Val di Pesa and Tavarnelle Val di Pesa, Tuscany.

It currently plays in Serie D.

Contents [hide]
1 History
1.1 Foundation
1.2 Serie D
2 External links

San Donato Tavarnelle

Full name	Associazione Sportiva Dilettantistica San Donato Tavarnelle
Founded	2006
Ground	Stadio San Donato in Poggio, Tavarnelle Val di Pesa, Italy
Capacity	400
League	Serie D
2013-14	Eccellenza Tuscany/B, 1st

History [edit]

Foundation [edit]

The club was founded in 2006 after the merger of *San Donato* and *Tavarnelle*.

Serie D [edit]

In the season 2013-14 the team was promoted for the first time, from Eccellenza Tuscany/B to Serie D.

External links [edit]

- Official homepage

V·T·E Serie D seasons [show]

V·T·E Football in Italy [show]

Figure 1: Visual structure of a typical Wikipedia page

4.2 Evaluation

For evaluation of ranked results, we utilize the NDCG@k (normalized discounted cumulative gain) score. For the rank parameter k of NDCG@k, we set two values: fixed value(1000) and relative value(top 1% of the total nodes), to evaluate both the top part and the overall ranking result(viewing the 1% top part as the sample).

NDCG@k score is defined as follows:

$$NDCG@k = \frac{DCG@k}{iDCG@k} \qquad (11)$$

Here, DCG@k is given below with generalized rank:

$$DCG@k = \sum_{i=0}^{k} \frac{g(i)}{log_2(i+1)} \qquad (12)$$

Here, k is the rank range, and i is the rank of a node.

Similarly, iDCG@k is calculated with the same formula with an ideal rank result.

4.2.1 Editorial Team Assessment of the WikiProject

Since the NDCG@k score requires an ideal ranking, here we use the manually assessment result of articles from Wikipedia Editorial Team Assessment as our golden standard [1]. There are totally 5694854 articles assessed and classified into 7 classes. To observe the rank result of articles of all quality level, the gain score $g(i)$ is set as follows:

Table1: gain score scheme for NDCG@k score

Quality level of article *i*	**Gain score *g(i)***
FA(Featured Article)	4

A(A-class article)
&
GA(Good Article)
	3

B(B-class article)	2
C(C-class article)	1
Start(Start-class article) & Stub(Stub-class article) & Unassessed	0

4.2.2NDCG@k(part) score

To observe the rank result of article of high quality, here we use another kind of relevance score scheme to calculate the NDCG@k(part) score. In this case, only the rank of articles of FA, GA, A quality level can influence the NDCG@k score. Here is the scheme:

Table 2: gain score scheme for NDCG@k(part) score

Quality level of article *i*	**Gain score *g*(*i*)**
FA(Featured Article)	4

A(A-class article)
&
GA(Good Article)
	3

B(B-class article)	0
C(C-class article)	0
Start(Start-class article) & Stub(Stub-class article) & Unassessed	0

4.3 Parameter setting

For the CWPR, we set γ taking values of {0.0, 0.2, 0.5, 0.7, 0.9, 1.0}. Here the CWPR(0) is actually the original PageRank and for the CWPR(1.0) only click counts affects the factor *f*;

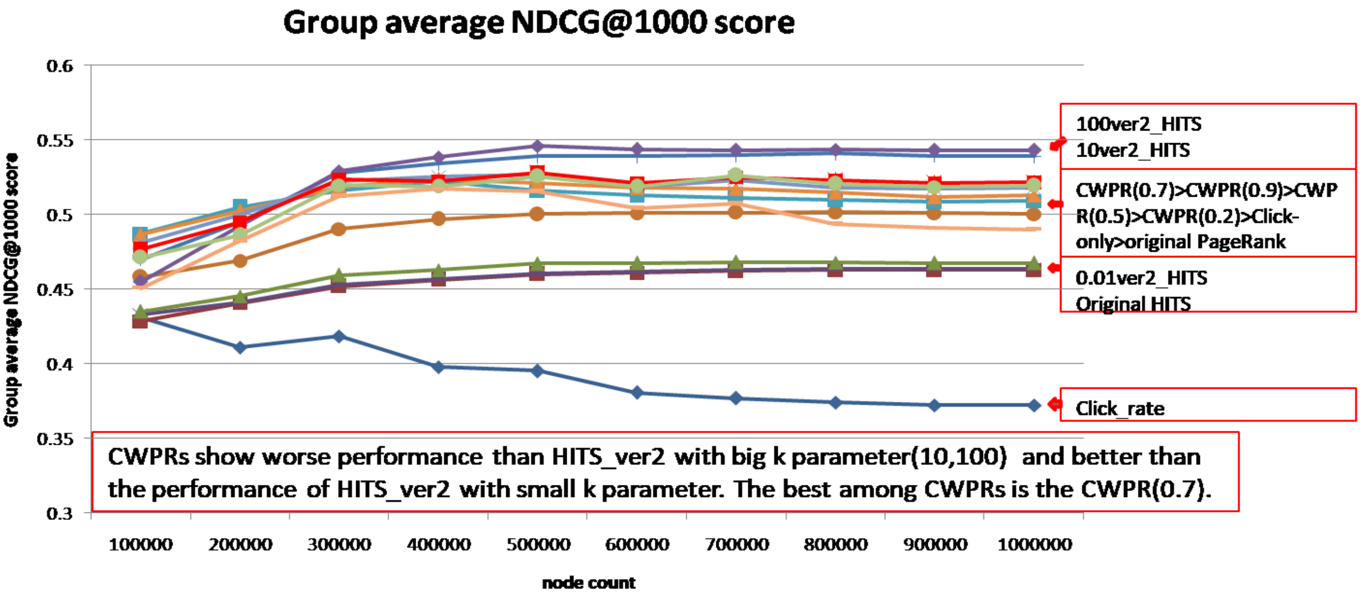
For the click count-weighted HITS ver2, we set its *k* taking values of {0.01, 0.1, 1, 10, 50, 100};

4.4 Result

We try different k values for the click count-weighted HITS ver2. Here, for easy to read, we only show the representative result graph with part of the methods.

4.4.1 NDCG@k score result

Since the NDCG@k score considers the rank of articles of all quality levels, this result in Figure2 shows the performance of the overall ranking result.



CWPRs show worse performance than HITS_ver2 with big k parameter(10,100) and better than the performance of HITS_ver2 with small k parameter. The best among CWPRs is the CWPR(0.7).

Figure2: Group Average NDCG@1000 score among different groups

The Figure2 is the group average NDCG@1000 score. Each node is the average NDCG score for one group by one method setting. From the Figure2, we can find: the HITS_ver2 algorithm with big parameter k(10,100) show better results than the CWPRs while the HITS_ver2 with small k(0.01,0.1,1) show worse results than the CWPRs.

Among CWPRs, the performance improves first as γ increases and get worse as γ becoming larger than 0.7. While CWPR(0), which is original PageRank, shows almost the worst result.

Also the group average NDCG score with k = 1% node count results are shown in Figure3.

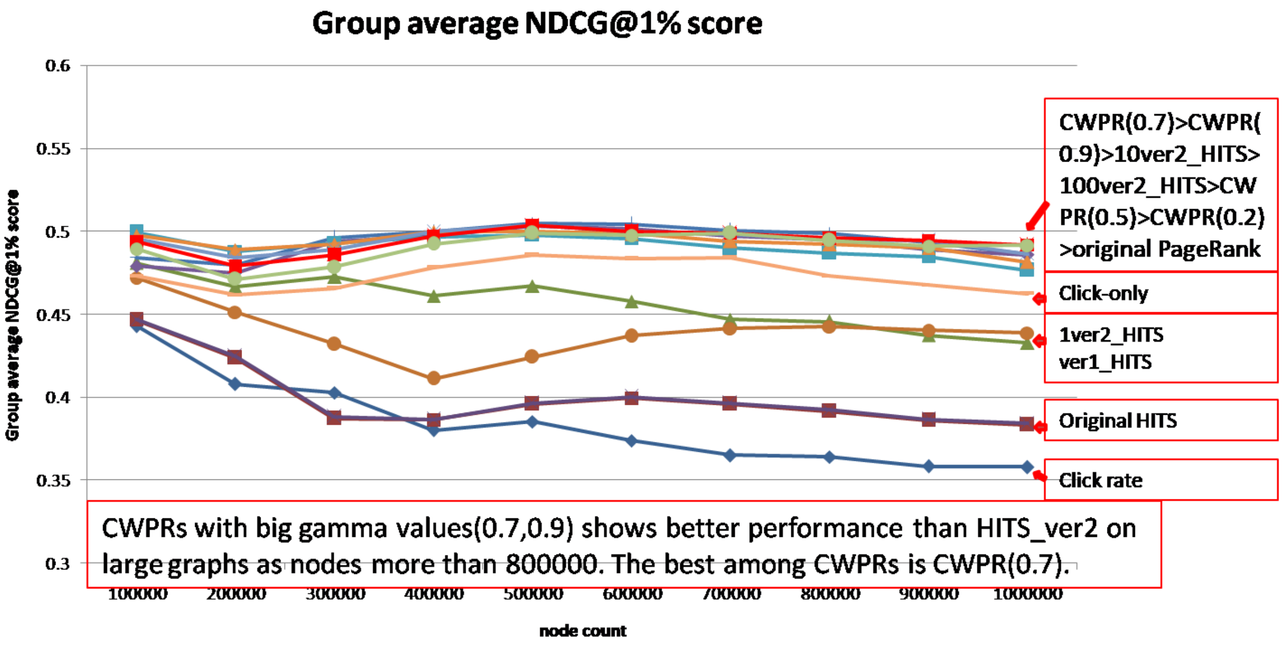


Figure3: Group Average NDCG@1% score among different groups

In Figure3, we can find the CWPR with big γ values(0.7, 0.9) have a lower decreasing tendency as the node count increases and shows better performance than other methods as the node count exceeds 800000, which means CWPR can achieve better results on large graph with big γ values(0.7, 0.9).

4.4.2 NDCG@k(part) score result

On the other hand, the NDCG@k(part) score only cares about the articles of high quality levels (FA,GA,A class), thus this result in Figure 4 and Figure 5 focus more on the performance of the ranking of high quality articles.

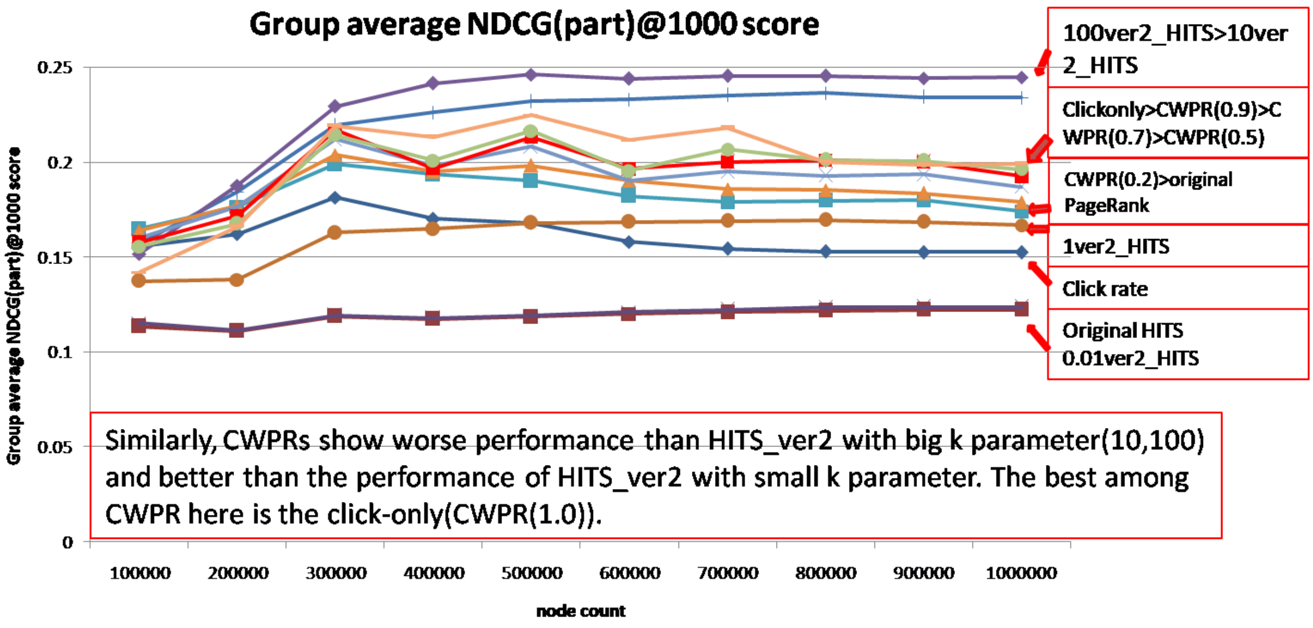


Figure 4: Group Average NDCG@1000(part)score among different groups

From the Figure4, we can find that the NDCG(part)@1000 scores of different methods are very close to each other. The top deep blue lines are the HITS_ver2 with big k parameters while the bottom purple lines are the results of the HITS_ver2 with small k parameters. The CWPRs are at the middle part of the figure. However, the performance among CWPRs with different γ values is a little different. The performance of CWPR improves as the γ increases.

The group average NDCG(part) score with k = 1% node count results are shown in Figure5.

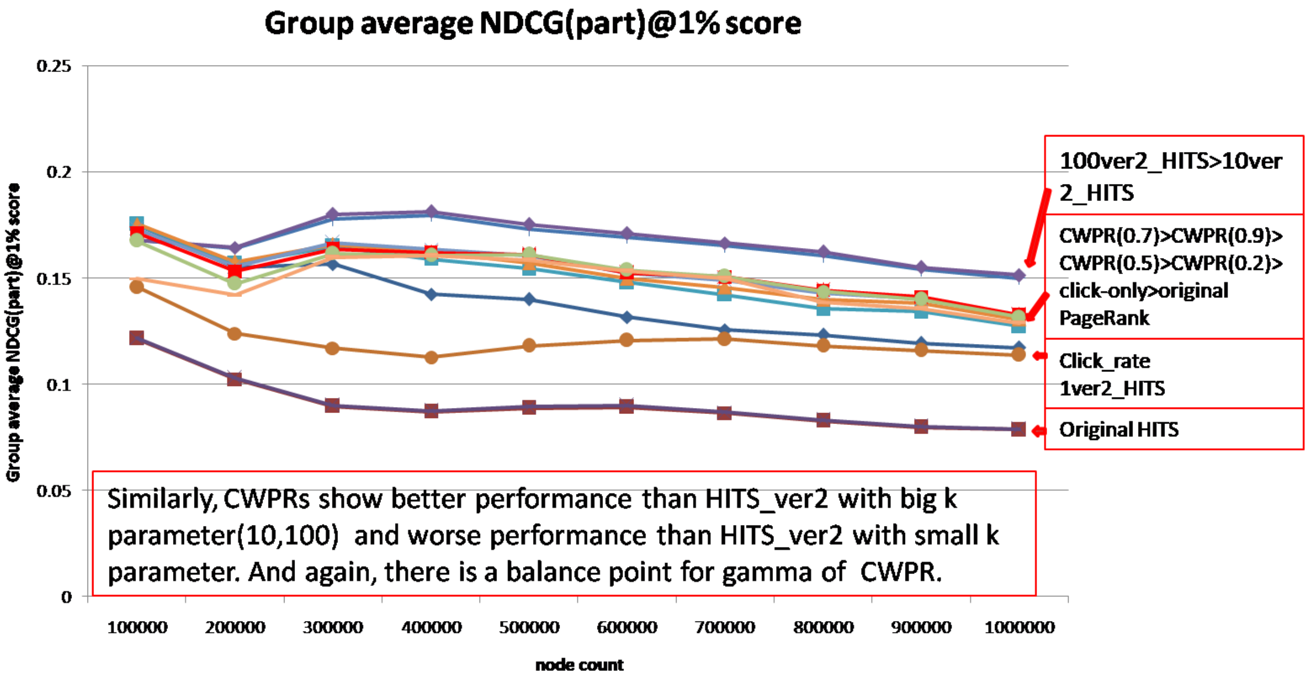


Figure 5:Group Average NDCG@1%(part) score among different groups

From the Figure5, we can find that although the CWPR with different γ values show little difference between each other and show worse performance than the HITS_ver2 with its parameter k = 10, 100. The CWPR perform better than all other cases. This means the CWPR show better stability and can almost get a good NDCG@1%(part) score.

4.5 Result Analysis

Combining Figure2 and Figure3, we can find that: for the top part, the CWPR can get good results but still a little worse than the HITS_ver2 with big k parameters. For the overall ranking, the CWPR with big γ values(0.7, 0.9) can outperform other methods as the graph becomes large. Also, from the performance among CWPR with different γ values, we can find there is a trade-off for the effect of link counts and click counts.

Combining the Figure4 and Figure5, we can find that: for the top part, the click counts have absolute advantage in finding high quality articles than the link counts. For the overall ranking, the trade-off between the effect of link counts and click counts still exists.

5. Conclusion &Future work

5.1 Conclusion

In this paper, we proposed a new algorithm called CWPR algorithm, which extends the original PageRank algorithm by integrating smoothing link counts and click counts. We use the English Wikipedia as our dataset and the manual quality assessment result from Editorial Team Assessment of WikiProject as our evaluation golden standard. And by NDCG@k score, we evaluate the result of ranking.

From the result of NDCG@k score and NDCG@k(part) score, we can find that the CWPR with bigγ values can always achieve good results and can outperform the original PageRank. It can also outperform other methods including the click count-weighted HITS algorithm in giving a good ranking of articles in English Wikipedia according to their quality, especially when the graph size becomes very large, which is often the case for web graph. So that suggests our proposed method is more suitable for large web graph ranking.

5.2 Future work

For future work, we plan to enlarge our dataset and integrate some other features relative to the content of the articles into our current methods which can also reflect the quality of articles.

References

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne,“Finding high-quality content in social

media," Proceedings of the 2008international conference on web search and data mining, ACM, February.2008, pp. 183-194

[2] J. E. Blumenstock, "Automatically Assessing the Quality of WikipediaArticles," unpublished.

[3] G. De la Calzada, and A. Dekhtyar, "On measuring the quality ofWikipedia articles," Proceedings of the 4th workshop on Informationcredibility, ACM, April. 2010, pp. 11-18.

[4] Dimitar Dimitrov, Philipp Singer, Florian Lemmerich, Markus Strohmaier, "Visual Positions of Links and Clicks on Wikipedia", Proceedings of the 25th International Conference Companion on World Wide Web WWW'16 Companion, April 11–15, 2016, http://dx.doi.org/10.1145/2872518.2889388.

[5] J. Kamps and M. Koolen, "Is wikipedia link structuredifferent?" Proceedings of the second ACM international conference on Web search and data mining, ACM, February, 2009, pp. 232-241.

[6] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," Journal of the ACM (JACM), 1999, 46(5), 1999, pp. 604-632.

[7] Amy N. Langville，Carl D. Meyer，James Hendler, "Google's PageRank and Beyond: The Science of Search Engine Rankings".

[8] R. Lempel，S. Moran,"The stochastic approach for link-structure analysis (SALSA) and the TKC effect", Computer Networks Volume 33, Issues 1–6, June 2000, pp. 387-401.

[9] Xuebo Liu, Shuang Ye, Xin Li, Yonghao Luo, Yanghui Rao "ZhihuRank: A Topic-Sensitive Expert Finding Algorithm in Community Question Answering Websites", ICWL 2015: Advances in Web-Based Learning -- ICWL 2015, pp. 165-173.

[10] N. Lipka and B. Stein, "Identifying featured articles in Wikipedia:writing style matters," Proc. 19th Int. Conf. World Wide Web, ACM,April. 2010, pp. 1147-1148.

[11] B. de La Robertie，Y. Pitarch，O. Teste, "Measuring Article Quality in Wikipedia using the Collaboration Network", Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015,pp. 464-471.

[12] Y. Suzuki and M. Yoshikawa, "Mutual evaluation of editors and textsfor assessing quality of Wikipedia articles," Proceedings of the EighthAnnual International Symposium on Wikis and Open Collaboration,ACM, Aug. 2012, pp. 18.

[13] M. Thelwall, ed, "Link analysis: An information science approach", Emerald Group Publishing Limited, 2004.

[14] E. Wulczyn and D. Taraborelli, Wikipedia clickstream. figshare, 2015,doi:10.6084/m9.figshare.1305770.

[15] W. Xing, A. Ghorbani, "Weighted PageRank algorithm", Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.

[16] Linfeng Yu and Mizuho Iwaihara, "Finding high quality documents through link and click graphs", Proc. 9th International Conference on E-Service and Knowledge Management (ESKM 2018), Yonago, July 2018 .

[17] https://dumps.wikimedia.org/other/clickstream/2018-05/clickstream-enwiki-2018-05.tsv.gz