# Quality Estimation of Local Contents Based on PageRank Values of Web Pages

Yutaka KABUTOYA
Informatics of the faculty of Engineering
Kyoto University
Yoshida-honmachi, Sakyo-ku
Kyoto, 606-8501 Japan
kabutoya@dl.kuis.kyoto-u.ac.jp

Takayuki YUMOTO Satoshi OYAMA Keishi TAJIMA Katsumi TANAKA
Graduate School of Informatics
Kyoto University
Yoshida-honmachi, Sakyo-ku
Kyoto, 606-8501 Japan
{yumoto, oyama, tajima, tanaka}@dl.kuis.kyoto-u.ac.jp

## Abstract

*Recently, it is getting more frequent to search not Web contents but local contents, e.g., by Google Desktop Search. Google succeeded in the Web search because of its PageRank algorithm for the ranking of the search results. PageRank estimates the quality of Web pages based on their popularity, which in turn is estimated by the number and the quality of pages referring to them through hyperlinks. This algorithm, however, is not applicable when we search local contents without link structure, such as text data. In this research, we propose a method to estimate the quality of local contents without link structure by using the PageRank values of Web contents similar to them. Based on this estimation, we can rank the desktop search results. Furthermore, this method enables us to search contents across different resources such as Web contents and local contents. In this paper, we applied this method to Web contents, calculated the scores that estimate their quality, and we compare them with their page quality scores by PageRank.*

## 1. INTRODUCTION

When we use web search engines, we get a huge amount of web search results. Those search results, however, include many web pages that are no use for some users, such as personal diary blogs. As a result, We seldom want all of the search results, and we usually use only highly ranked ones. Therefore, ranking of the Web search results is very important and it is ideal that all the highly ranked Web pages are useful.

Google succeeded because of its "PageRank" algorithm. "PageRank" algorithm is based on a recursive relation "a page linked from many good quality pages also has good quality". Google estimates the quality of each web page by this algorithm.

On the other hand, local search services, such as Google desktop search, gets more and more popular. Ranking was not important for local search system before because we had relatively a small amout of contents on personal computers. Recently, however, the cheaper large capacity storages gets, the larger amount of local contents we have. In this paper, we call the contents stored in personal devices, such as PCs, DVD-Recorders or HD-Recorder, not in web "local contents".

"PageRank" algorithm, which made Google succeed, is based on the hyperlink structure between web pages. Therefore, we cannot apply it to local contents which do not have the hyperlink structure. In this research, we hypothesize "the more similar contents are, the more similar quality they have". Based on this hypothesis, we use the PageRank values of similar Web pages in order to estimate the quality of local contents. Therefore, we call the quality of local contents estimated by this method "virtual PageRank".


Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)
0-7695-2571-7/06 $20.00 © 2006 IEEE

## 2. CALCULATION OF VIRTUAL PAGERANK

In this section, we explain how we calculate the Virtual PageRank scores of local contents.

### 2.1. Assumption

We assume that a page linked from many good quality pages also has good quality. The reason is shown as follows. The goal of PageRank algorithm is to exclude contents which is no use for some people, such as diary cites, from the higher ranks in the search results. As a consequence of that, highly ranked pages in Google search results are similar to each other, and low ranked pages are also similar to each other.

### 2.2. Calculation

Let $W_i (i = 1, 2, \ldots, N)$ be the $i$th page in the ranking of a Google search result, and let $L$ be a local content. $sim(W_i, L)$ is cosine similarity between $W_i$ and $L$. Let $PR(Q, C)$ be the real PageRank score of a content $C$ when query $Q$ is provided.

Then, the virtual PageRank score of a local content $L$ under $Q$, which we denote by $PR'(Q, L)$, is defined as follows:

$$PR'(Q, L) = \frac{\sum_i sim(W_i, L) \cdot PR(Q, W_i)}{\sum_i sim(W_i, L)} \quad (1)$$

On the other hand, we also define Virtual PageRank for Web page $W_i$ as follows:

$$PR'(Q, W_i) = \frac{\sum_{j \neq i} sim(W_j, W_i) \cdot PR(Q, W_j)}{\sum_{j \neq i} sim(W_j, W_i)} \quad (2)$$

### 2.3. Normalization

In order to discuss the validity of our method, we have to compare with Virtual PageRank scores with the "real" PageRank scores. For such a comparison, we need to normalize the virtual PageRank scores along to the "real" PageRank scores. [2] proposed 3 normalization schemes as follows.

1. **Standard**
   If the maximum value and the minimum value of the scores are same, it is possible to compare them with each other.
2. **Sum**
   If the minimum value and the sum of the scores are same, it is possible to compare them with each other.
3. **ZMUV**
   If the mean and the variance of the scores are same, it is possible to compare them with each other.

In this paper, in order to compare the virtual PageRank score with the real PageRank score, we normalize the virtual PageRank score by the **Standard** scheme.

Let $min$ be the minimum value of $PR(Q, W_i)$, let $max$ be the maximum value of $PR(Q, W_i)$, let $min'$ be the minimum value of $PR'(Q, W_i)$, and let $max'$ be the maximum value of $PR'(Q, W_i)$. Then normalized Virtual PageRank score $PR''$ is define by the equation below:

$$PR''(Q, C) = \frac{max - min}{max' - min'}(PR'(Q, C) - min') + min \quad (3)$$

(4) and (5) follow from Eq. (3).

$$\min_i(PR''(Q, W_i)) = \min_i(PR(Q, W_i)) \quad (4)$$

$$\max_i(PR''(Q, W_i)) = \max_i(PR(Q, W_i)) \quad (5)$$

## 3. EXPERIMENT FOR NORMALIZATION

We run some experimentes in order to get the values of $min'$ and $max'$.

### 3.1. Apprximation of the PageRank Score

Because we cannot konw the exact values of the PageRank scores, we approximate the "real" PageRank scores by the Eq. (6).

$$S(Q, W_i) = N - i + 1 \quad (6)$$

(7), (8) follows from Eq. (6).

$$min = 1 \quad (7)$$

$$max = N \quad (8)$$

### 3.2. The Procedure of the Experiment

We shall now describe the experimental procedure.

1. Create 10 queries $Q_k (k = 1, 2, \ldots 10)$.
2. Get 1st-100th pages of Google search result $W_i (i = 1, 2, \ldots, 100)$.
3. calculate $PR'(Q_k, W_i)$ by (2)
4. Get the minimum value and maximum value of $PR'$

Altogether, we apply Virtual PageRank to each web page of Google search results.

(7) and (8) make us suspect that $min'$ and $max'$ are related to the number of Web pages $N$. Then we run the following experiments.


Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)
0-7695-2571-7/06 $20.00 © 2006 IEEE

**Table 1. Terms of Queries**

| Query | Terms |
|---|---|
| $Q_1$ | Kyoto University (in Japanese) |
| $Q_2$ | Johjima Mariners (in Japanese) |
| $Q_3$ | Google PageRank |
| $Q_4$ | Firefox |
| $Q_5$ | ShihTzu (in Japanese) |
| $Q_6$ | Hokkaido Soupcurry (in Japanese) |
| $Q_7$ | Java |
| $Q_8$ | News23 (in Japanese) |
| $Q_9$ | iPod |

**Table 2.** $min'$ **and** $max'$ **for queries** $Q_k$

| Query | $min'_1$ | $max'_1$ |
|---|---|---|
| $Q_1$ | 32.153 | 68.395 |
| $Q_2$ | 34.939 | 69.431 |
| $Q_3$ | 36.407 | 69.855 |
| $Q_4$ | 35.239 | 81.430 |
| $Q_5$ | 34.278 | 64.200 |
| $Q_6$ | 39.582 | 59.896 |
| $Q_7$ | 39.449 | 59.644 |
| $Q_8$ | 32.313 | 65.428 |
| $Q_9$ | 21.000 | 63.819 |

1. Select one query from queries $Q_k(k = 1, 2, \ldots, 10)$.
2. Calculate $min'$ and $max'$ when $N$ is 50, 60, 70, 80, 90, 120, 150, and 200.

### 3.3. Result of Experiments

We shall show the values of $min'$ and $max$ for queries $Q_k$ in table2.

The minimum value and the maximum value of the "real" PageRank scores is respectively 1 and 100. As table 2 shows, the values of $min'$ and $max'$ looks like to be independent of a query $Q_k$.

Graph 1 shows the relation between the number of Web pages $N$ and $min'$, $max'$, where the query is $Q_1$.

Graph 1 tells us that both $min'$ and $max'$ are proportional to $N$, hence the below equation follows.

$$min' = 0.35N \tag{9}$$

$$max' = 0.65N \tag{10}$$

The below equation follows from (3), (9) and (10).

$$PR''(Q, W_i) = \frac{10}{3}(PR'_1(Q, W_i) - 0.35N) + 1 \tag{11}$$

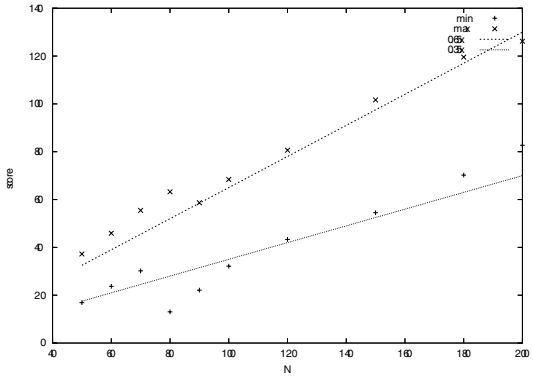


**Figure 1. relation between** $N$ **and** $min'$, $max'$

**Table 3.** $r$**: correlation between** $PR$ **and** $PR''$

| Query | $r$ | Query | $r$ |
|---|---|---|---|
| $Q_1$ | 0.386 | $Q_6$ | 0.614 |
| $Q_2$ | 0.595 | $Q_7$ | 0.040 |
| $Q_3$ | 0.403 | $Q_8$ | -0.108 |
| $Q_4$ | 0.689 | $Q_9$ | 0.337 |
| $Q_5$ | 0.339 | | |

## 4. COMPARISON BETWEEN PAGERANK AND VIRTUAL PAGERANK

### 4.1. Experiment

We divided Web pages of Google search results $W_i(i = 1, 2, \ldots, 100)$ into two groups. There are 30 web pages in the one group (group A), and 70 web pages in the other group (group B). We calculated the Virtual PageRank score of each page in the group A from PageRank scores of the pages in the group B. Then, we compared Virtual PageRank scores with "real" PageRank scores of the pages in the group A.

Let $W_{a_p}(p = 1, 2, \ldots, 30)$ denote a Web page of the group A. Table 3 shows correlation $r$ between $PR''(Q_k, W_{a_p})$ and $PR(Q_k, W_{a_p})$. Fig. 2 shows relation between "real" PageRank scores and Virtual PageRank scores where the query is $Q_5$, and Fig. 3 shows where the query is $Q_9$.

In this experiments, the web pages whose Virtual PageRank scores are quite different from those PageRank scores must be analyzed.


Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)
0-7695-2571-7/06 $20.00 © 2006 IEEE

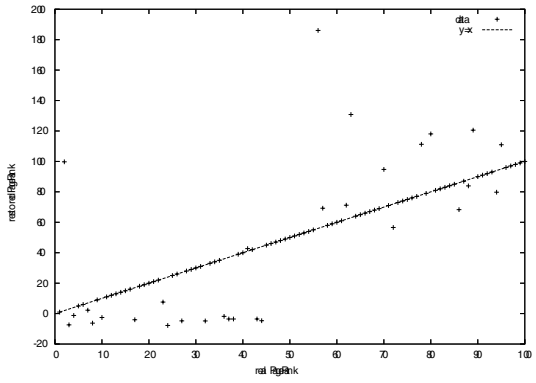


**Figure 2. relation between $PR$ and $PR''$ where the query is $Q_5$**

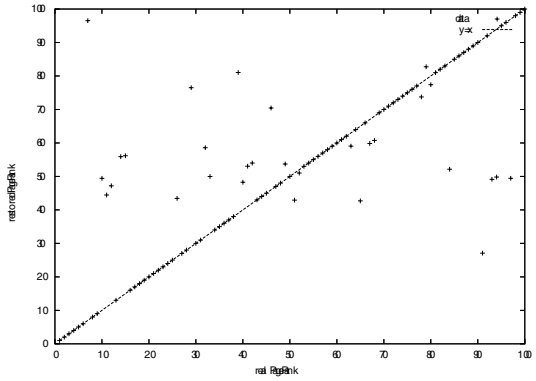


**Figure 3. relation between $PR$ and $PR''$ where the query is $Q_9$**

### 4.2. Consideration

Virtual PageRank may enable us to clarify whether a content is in the high rank or in the low rank of the Google search results, for example, where the query is $Q_5$. On the other hand, where the query is $Q_9$, there are no correlation between PageRank scores and Virtual PageRank scores. Hence, whether Virtual PageRank is effective or not is dependent on the query.

## 5. CONCLUSION

In this research, we proposed a method to estimate the quality of local contents which do not have the hyperlink structure from the PageRank score of each page of Google search results. Then, we conducted the experiment to apply this method to Web pages and to compare Virtual PageRank scores with PageRank scores. This experiment yields that there is possibility that this method makes us easy to find out useful local contents.

Our future work is shown as follws.

- PageRank algorithm is a method to estimate the quality of Web pages from the aspect of popularity. Similarly, we should propose a method to estimate the popularity of local contents.
- In this paper, Virtual PageRank scores are normalized according to **Standard** scheme. It wiil be interesting to try other method of normalization.
- It is necessary to evaluate Virtual PageRank algorithm.

## ACKNOWLEDGMENTS

Our work was supported in part by the Japanese Ministry of Education, Culture, Sports, Science, and Technology under a Grant-in-Aid for Software Technologies for Search and Integration across Heterogeneous-Media Archives. It was also supproted by a Special Research Area Grant-In-Aid for Scientific Research, in 2005, under a project titled Research for New Search Service Methods Based on the Web's Semantic Structure(Project No. 16016247; Representative, Katsumi Tanaka), and it was supported by the Informatics Research Center for Development of Knowledge Society Infrastructure (COE Program of the Japanese Ministry of Education, Culture, Sports, Science, and Technology).

## References

[1] Google JAPAN
`http://www.google.co.jp/`.

[2] M. Montague and J. Aslam. Relevance score normalization for metasearch. *Proc. 10th ACM International Conference on Information and Knowledge Management*, pages 427–433, 2001.

[3] L. Page, S. Brin, R. Montwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. *Stanford Digital Libraries Working Paper*, 1998.


Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)
0-7695-2571-7/06 $20.00 © 2006 IEEE