Weighted PageRank Algorithm to Consider the Inter-Page Similarity

Sang-Yeon Lee, Young-Gi Kim, Keon-Myung Lee
Dept. of Computer Science, Chungbuk National University,Korea
E-mail : {jaeimveilion, yogi101,kmlee}@chungbuk.ac.kr

1. Introduction

Information retrieval from huge data repositories, such as Web and big data stores, might provide overwhelmingly many records or pages. It is important to show handful of retrieved data to users early on because it is another burdensome to search for interesting ones from large amount of retrieved results. Ranking the retrieved results has had been actively studied in web search literature. The well-known algorithms include PageRank, HITS, and SALSA and many of their variants have been proposed.[1-4]Those algorithms make use of link structure of Web to determine the ranks of pages. PageRank assumes that surfers walk randomly over the web and tries to determine the stationary distribution of the surfers. It basically treats a surfer on a page to have the same probability to the neighboring pages. This treatment makes it easy to compute the stationary distribution, but it does not reflect the actual behaviors of surfers. To provide different transition probability between pages, several weighted PageRank algorithms have been proposed which use the information of topological structures. Surfers may well as prefer to follow the neighboring pages which might have similar contents to the current page. From this observation, we propose a weighted PageRank algorithm in which transition probability is proportional to the similarity between neighboring pages.

2. Related Works

2.1 PageRank algorithm

PageRank is the frontier algorithm which ranks pages according to the link structure of Web, in which pages are regarded as nodes and hyperlinks as edges of a graph. Each node has its own Rank value and distributes it evenly to its neighbors. The stationary distribution of the Ranks is considered as the final Rank scores of pages. The retrieval pages are basically sorted according to their Rank scores. The Rank score r_j of node *j* is computed as follows:

$$r_j = \sum_{i \to j} \frac{r_i}{d_{out}(i)} \quad (1)$$

$$\sum r_j = 1 \quad (2)$$

where $d_{out}(i)$ denotes the number of out link of node *i*. The updates are repeated until no meaningful changes are observed.

2.2 Weighted PageRank Algorithm

Surfers do not make random walks as in PageRank. Some variants, called weighted PageRank, have been tried to make uneven probabilistic transitions to neighboring pages. Xing et al.[4] determines the weights based on the number of in-links of neighboring pages, in a way that the more in-links neighboring page, the larger weight it has. Kumar et al.[5] proposed a weighted PageRank algorithm which gives more Rank value to the outgoing links that are more visited by users.

3. The Proposal Algorithm

The proposed algorithm is a weighted PageRank algorithm in which weights are determined based on the similarity between neighboring pages. As the neighboring pages have higher similarity, the weight on the corresponding edge gets higher. The weight w_{ij} on the edge from node *i* to *j* is determined as follows:

$$w_{ij} = \frac{s_{ij}}{\sum_{k \in N(i)} s_{ik}} \quad (3)$$

where s_{ij} is the similarity between pages *i* and *j* and $N(i)$ denotes the pages to which page *i* is pointing.

The Rank r_j of page *i* is updated as follows until the Ranks converge:

$$r_j = \sum_{i \in I(j)} \beta w_{ij} r_i + (1-\beta)\frac{1}{N} \quad (4)$$

where β indicates the rate for the teleportation as in PageRank, $I(j)$ is the pages which point to page *j*, and *N* is the total number of pages.

In order to compute the similarity of two pages, the proposed method uses the following method. First, it chooses keywords for each page. The keywords in a pages are determined using the TFIDF technique. TF(term frequency) is the frequency of a word in a document, and IDF (inverse document frequency) is the reciprocal of DF (document

frequency) which is the number of pages containing the word. For a word in a document, its TFIDF is computed as follows:

$$\text{TFIDF} = \frac{TF}{\log\left(\frac{N}{DF}\right)} \tag{5}$$

For each page, the keywords are selected on basis of their TFIDF, in which higher TFIDF indicates higher potential to be keyword.

Then the similarity s_{ij} between pages i and j is computed using the cosine distance for the corresponding keyword vectors K_i and K_j.

$$s_{ij} = \frac{K_i \cdot K_j}{|K_i||K_j|} \tag{6}$$

4. Experiments

The proposed method has been implemented and tested on Korean Wikipedia web pages. Figure 1 shows the experiment system architecture. The entire pages from ko.wikipedia.org were crawled and parsed using a Korean morphological analyzer to extract the stem words. Then, the TFIDFs for the stem words of each page were computed to determine keywords, and a keyword vector was constructed for each page. The similarities between neighboring pages were computed using the cosine distance. After that , the proposed weighted PageRank algorithm were applied to the graph structure of Korean Wikipedia along with the similarity information to determine the Rank score.

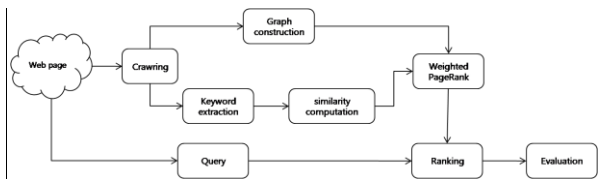

Figure 1. Experiment System Architecture

In Korean Wikipedia, the number of unique crawled pages was about 300,000 and the system was implemented using MapReduce paradigm on Hadoop platform. For the purpose of comparative studies, the conventional PageRank algorithm was also implemented. The performance of the proposed method was evaluated using the performance metric NDCG (normalized discounted cumulative gain) [3]. In the performance evaluation, the relevance of pages to query was evaluated at four levels, i.e., *perfect*(5), *good*(3), *fair*(1), and *bad*(0). A set of queries were collected for the experiments and the pages were searched up to one hundred for each query based on keyword-query match. The relevance of retrieved pages to queries was manually evaluated. After that, the retrieved pages were sorted according to PageRank's rank and the proposed weighted PageRank, respectively. For the sorted sequence, the NDCG were computed.

5. Conclusion

Ranking of retrieved results is one crucial issue in information retrieval system. This paper proposed a weighted PageRank algorithm which takes into account the similarity between linked pages. The similarity is evaluated based on the keywords of the pages. Due to volume of data, the experiment system was implemented using MapReduce paradigm on Hadoop platform. The experiments showed interesting results although it goes the extra mile to get similarity information between neighboring pages.

6. Acknowledgement

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program(NIPA-2013-H0301-13-4009) supervised by the NIPA(National IT Industry Promotion Agency)

7. References

[1] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", Stanford Digital Libraries Working Paper, 1999.
[2] K. Kumar, Abhaya, F. D. Mukoko, "PageRank algorithm and its variations: A survey report," ISOR J. of Comp. Eng.,Vol.14, I.1, pp.38-45, 2013.
[3] M. Najork, "Comparing the Effectiveness of HITS and SALSA", Proc. of CIKM'07, 2007.
[4] N.Duhan, A. K. Sharma, and K. K. Bhatia, "Page Ranking Algorithms: A Survey", proc. IEEE International Advance Computing Conference (IACC 2009)
[5] G. Kumar and N. Duhan, and A. K. Sharma, "Page Ranking Based on Number of Visits of Links of Web Page", proc. India International Conference on Computer & Communication Technology (ICCCT), 2011