韓國 컴퓨터情報學會 論文誌
第10卷 第1號, 2005. 3.

2005-3-1-1-8

하이퍼텍스트 정보 관점에서 의도적으로 왜곡된 웹 페이지의 검출에 관한 연구

이 우 기*

Detecting Intentionally Biased Web Pages In terms of Hypertext Information

Woo-Key Lee*

요 약

웹(World Wide Web)은 정보의 저장 및 검색에 있어서 보편적인 매체가 되고 있다. 웹에서는 일반적으로 검색엔진(Web search engine)을 통해 정보 검색을 수행하지만, 그 결과가 사용자의 요구와 늘 일치하는 것은 아니며 때로는 의도적으로 조작된 검색 결과가 제시되기도 한다. 웹 페이지에 대한 평가를 조작하는 것을 의도적 조작이라고 부른다. 최근에 가장 각광을 받는 링크 기반 검색 방식에는 의도적 조작이 상대적으로 어렵지만, 링크 기반 검색 방식의 대표격인 구글의 페이지 점수법(PageRank algorithm)도 구글밤처럼 조작할 수 있는 방법이 있다. 본 논문에서는 기본적으로 링크 기반 검색 방식을 기초로 웹을 하나의 유향 그래프(directed graph)로 인식하여 각 웹 페이지들은 하나의 노드로, 하이퍼텍스트 링크를 에지(edge)로 표현하며, 하이퍼텍스트 정보관점에서 링크 내역과 대상 페이지(target page) 사이의 유사도(similarity)를 구하고, 이것을 이용하여 페이지 점수화 (PageRank) 접근법의 전이 행렬(transition matrix)을 재구성하는 방법을 취했다. 결과적으로 기존의 점수화 방법과 비교하여 효과가 60% 이상 될 수 있음을 입증했다.

Abstract

The organization of the web is progressively more being used to improve search and analysis of information on the web as a large collection of heterogeneous documents. Most people begin at a Web search engine to find information, but the user's pertinent search results are often greatly diluted by irrelevant data or sometimes appear on target but still mislead the user in an unwanted direction. One of the intentional, sometimes vicious manipulations of Web databases is a intentionally biased web page like Google bombing that is based on the PageRank algorithm, one of many Web structuring techniques. In this thesis, we regard the World Wide Web as a directed labeled graph that Web pages represent nodes and link edges. In the present work, we define the label of an edge as having a link context and a similarity measure between link context and target page. With this similarity, we can modify the transition matrix of the PageRank algorithm. By suggesting a motivating example, it is explained how our proposed algorithm can filter the Web intentionally biased web pages effective about 60% rather than the conventional PageRank.

▸ Keyword : Search Engine, Information Retrieval, Web Spamming, PageRank, Transition Matr

• 제1저자 : 이우기
• 접수일 : 2005.01.17, 심사완료일 : 2005.03.17
* 한국재활복지대학 정보보안과 부교수, ** 강남대학교 전자시스템공학부 교수

Ⅰ. 서 론

The organization of the web is progressively more being used to improve search and analysis of information on the web as a large collection of heterogeneous documents. The web spam refers to hyperlinked pages on the web that are created with the intention of misleading search engines. Traditional search engines based on the information retrieval techniques are well known to weaknesses for term spamming including body spam, title spam, meta tag spam, anchor text spam, and url spamso that the added keywords can be invisible to persons through ingenious use of color representations, but can mislead the search engines. Another web spamming technique is the creation of a large number of fake web pages, all directingto a target page. Since many search engines take into account the number of incominglinks in ranking pages, the rank of the target page is likely to increase, and appear earlier in query result sets. We call this spam pages as an intentionally biased web page that can use various techniques to achieve higher -than-deserved rankings in a search engine's results [1]. A Web structuring technique in terms of the number of incoming links is widely put to use and is expected to minimize these weakness[2]. However, the results of many Web search engines using Web mining techniques are equally hard to assay since search engines usually return huge lists of URLs, most of which can be judged almost irrelevant to the query[3]. For example, the rank can easily be manipulated by generating of a large number of bogus web pages and then all the pages point a single target web page. In identifying the reason for the weaknesses we can look to the inaccuracy of the Web mining algorithm on one hand, and Web pages that are deliberately composed to spam the search engine, on the other. Like traditional Information retrieval techniques, Web content mining alerts the discovery of useful information on the basis of match percentages gathered by scanning Web contents, related data, and uploaded documents [2, 14]. Yahoo, DMOZ, and many other Web search engines use this type of algorithm. However, there are two main reasons why traditional information retrieval (IR) techniques may not be effective enough in ranking query results.

In this thesis, we have the following assumptions: First, the Web is a vast ocean of information, with great variation in the amount, quality, and type of information contained in Web pages. Thus, many pages that contain the search terms may be of poor quality or not be relevant. Second, many Web pages are not sufficiently self-descriptive, so the IR techniques that examine the contents of a page alone may not work well. Though the link structure of the Web contains important implied information, and can help in filtering or ranking Web pages[4], and while there are several Web search engines using Web structuring techniques, Google being the most popular example, Web structuring techniques cannot solve the problem perfectly because they still have several weak points. A chief observation is that Web pages are frequently manipulated by adding misleading terms, so that search engines rank them higher(spamming). Hence, techniques that base their decisions on the content of pages alone are easy to manipulate. Consequently, some Web pages are erroneously ranked higher than others in spite of their content deficiency. For instance, Google bombing[1, 5] uses this weak point of the PageRank algorithm.

In this thesis, we will recognize the weak points of the Web structuring technique and suggest an alternative for that. With this algorithm, Web search engines can filter the intentionally biased web pages effectively and offer the correct information to users.This work is organized as follows. Section

two presents a review of Web structuring and Hypertext information. In section three, we define the problem that is focused upon and will be solved in this thesis and constitutes the similarity measures. Section four describes the performance analysis to solve the link based structuring problem and we work through a brief example to show how the algorithm can solve the problem. Then we put forth some concluding remarks and suggestions for the future works.

II. Hypertext Information

The web can be viewed as a digraph consisting of a set of web pages[1] or of web sites[2]. A web site, specifically, has an initial node called homepage and usually many other nodes as web pages. Then the WWW can be viewed as a hierarchy of web objects. The web-as-a-graph approach can be a starting point to generate a structure that can be used for web site designers, search engines, web crawling robots, and web marketers and analysts [14]. The graph structure per se, however, is too complex to view in one shot since the whole web is so enormously interconnected.

There are two major link-based search algorithms, HITS (Hypertext Induced Topic Search) and PageRank. The basic idea of the HITS algorithm is to identify a small sub-graph of the Web and apply link analysis on this sub-graph to locate the authorities and hubs for the given query. The sub-graph that is chosen depends on the user query. The selections of a small sub-graph (typically a few thousand pages), not only focus the link analysis on the most relevant part of the Web, but also reduce the amount of work for the next phase. The main weaknesses of HITS are known to non-uniqueness and nil-weighting [5]. THESUS suggested a domain based PageRank algorithm, but its limitation depends on the usefulness of the ontology and the thesaurus that the system tries to include semantics among Web documents. The link structure of the Web contains important implied information, and can help in filtering or ranking Web pages. In particular, a link from page A to page B can be considered a recommendation of page B by the author of A. Some new algorithms have been proposed that exploit this link structurenot only for keyword searching, but other tasks like automatically building a Yahoo-like hierarchy or identifying communities on the Web. The qualitative performance of these algorithms is generally better than the IR algorithms since they make use of more information than just the contents of the pages. While it is indeed possible to influence the link structure of the Web locally, it is quite hard to do so at a global level. So link analysis algorithms that work at a global level possess relatively robust defenses against spamming[1]. We will see the representative approaches such as Google Method, Hyperlink Information Method, and Singular Value Decomposition in more detail as follows.

The basic idea of PageRank[7] is that, if source page Ni has a link to target page Nj, then the author of source page Ni is conferring some importance to page Nj. Let O(Ni)be the out-degree of page Ni and let PR(Ni) represents the importance of page Ni. Then, the link (i, j) confers a certain number of units of rank to Nj. This simple idea leads to the following iterative fix-point computation that yields the rank vector $\overrightarrow{Rank}$ over all of the pages on the Web. If S is the number of pages, assign all pages the initial value 1/S. Let O(Ni) represent the number of links out of web node Ni. In the iteration, links between Web pages propagate the ranks. We choose a parameter d such that $0 < d < 1$; a typical value of d might lie in the range 0:1 < d < 0:15. Let pages Ni,, Ni link to page p. Let PR(Ni) be the PageRank of Ni and O(Ni) be the number of links out of Ni. Then the PageRank PR(Ni) of page Ni is defined to satisfy:

$$\forall i, \quad PR(N) = (1-d)\sum_{i=1}^{n} PR(N_i)/O(N_i) + d/S$$

The PageRank distribution has a simple interpretation in terms of a random walk. Imagine a web surfer who wanders the web. If the surfer visits page p, the random walk is in state p. At each step, the web surfer either jumps to a page on the web chosen uniformly at random, or the web surfer follows a link chosen uniformly at random from those on the current page. The former occurs with probability d, the latter with probability 1. Guaranteeing the rank vector to converge, PageRank algorithm uses the following equation with a damping factor (d).

In the previous section, we have already emphasized the importance of the link structure of the WWW. Nevertheless, we are not satisfied with the role of a hyperlink as a simple linkage between Web pages. Links themselves can have meanings or semantics. Many other works [3, 8, 9] have been produced from the viewpoint of link information and, in this thesis, we are operating from the perspective that links also have semantics. The issue of extracting keywords from links is important in the context of this thesis.

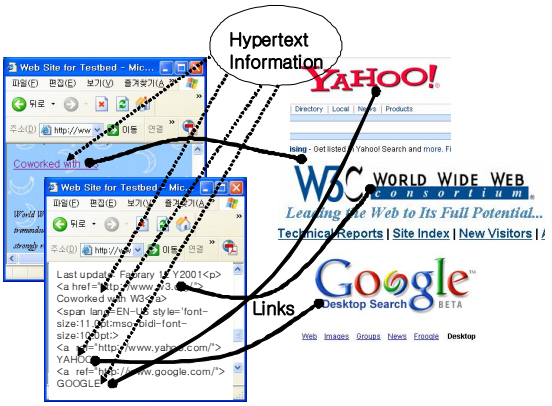

Fig. 1. Hypertext Information

In the work by Phelps and Wilensky[10], the idea of robust hyperlinks is introduced. Robust hyperlinks are considered to be those that contain descriptive information on the target document. According to the authors, this information can be limited to five words and empirical results are used to prove this. The lexical signature can be regarded as set of terms thatdescribes the target page of the link[11]. The Web is a large collection of heterogeneous documents. Web pages can contain multimedia such as images, sounds, flash, active-x, etc. and links to other documents through hypertext information. Hypertext links are being actively used to improve web search engine ranking, and enhance the web crawlers. The basic assumption in the hypertext analysis is that a link is created because of a weight endowment between the original document and the linked document. Hypertext has been utilized by Google to improve web search. PageRank algorithm by Google allows pages to be returned based on keywords occurring in inlink hypertext such as returning http://www.w3c.org/ for a query of web standards. The hypertext allows for linking words to destination pages, as shown in Figure 1.

III. Similarity Measure

In this paper, we use the randomized weight onto the Hypertext Information that can be one of the measures for the degree of co-occurrence between a link context and a sentence of the target page[17]. The degree of co-occurrence can be calculated with the following equation. This recursive definition gives each page a fraction of the rank of each page pointing to it inversely weighted by the number of links out of that page. Like other similarity measures such as term frequency and inverse of document frequency[16], cosine function[9], etc., if we consider all the terms of the target page, the similarity

must be influenced either by the length of the target page or by other terms in the target page. The similarity between query and sentences in the document can significantlyimprove the document retrieval effectiveness[9]. But, it is possible that some page is regarded as an intentionally biased web page although it is very relevant to the link context. Using the maximum value of similarity between the link context and each sentence in the target page, we prevent such misclassifications. In addition, we can consider the semantic of link context and page.

We can normalize the weight values of the link. It is the way to guarantee the convergence of the sum of the elements of the rank vector. If the sum of a row of the transition matrix is larger than 1, the sum of the elements of the rank vector will diverge. In contrast, if the sum is less than 1, the sum of the elements of the rank vector converges to 0. Thus, we must normalize the weight values to set the sum of a row of transition matrix to 1 [6, 8].

We keep the iteration of calculating the rank vector until the stop condition is true. There are two ways to stop calculating the rank vector, one is to stop when the number ofiterations equals to predefined value. And the other is to stop when the vector converges to the specific value, that is, the difference between ith rank vector and i+1th rank vector is less than the predefined threshold value.

One caveat is that the convergence of PageRank is only guaranteed if transition matrix T is irreducible (i.e. graph is strongly connected) and aperiodic [13]. The latter is guaranteed in practice for the Web, while the former is true only if all of the nodesin the graph are accessible; hence, we add a complete set of outgoing edges to nodes in G with out-degree 0 and damp the rank propagation by a factor of $(1-d)$ by adding a complete set of outgoing edges, with the equal weight one divided by the number of nodes to all the links [6]. In the calculation of rank vector, we use the damping factor as Google's PageRank algorithm and set the value to 0.85 [3, 4].

IV. Performance Analysis

(Figure 2) represents a motivating example to explain how the proposed algorithm finds the spamming link and reduces the importance of the intentionally biased web page. (Figure 2) shows an example graph with 8 pages and the corresponding links. We assume that Web page N8 is an intentionally biased web page and then links (1, 8), (4, 8), (6, 8) and (7, 8) would be the spamming links. The goal can be how the PageRank algorithm and our approach work on these kinds of intentionally biased web pages and links.

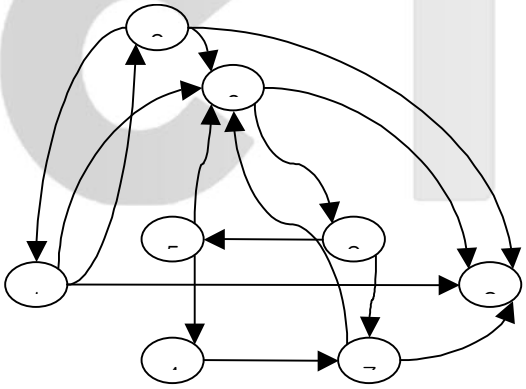

Fig 2. Example graph having 7 Web nodes

<Table 1> shows the information of each page including title, link and initial rank. Page 8 has no outgoing link and only incoming link from page 1, 2, 4, 6 and 7. Like simple example, initial rank value of each page was set to 1for preventing page rank from being influenced by initial value.

Table 1. Description of the Web Node and Links

Page	Links	Initial rank
1	(1,2), (1,6), (1,8)	1
2	(2,1),(2,6),(2,8)	1
3	(3,5),(3,7)	1
4	(4,7)	1
5	(5,4),(5,6)	1
6	(6,3),(6,8)	1
7	(7,6),(7,8)	1
8	No links	1

<Table 2> shows the link context including link string and link signature and weight value which was calculated with link context and target page. Lexical signature of the each link was composed with the keywords in the target page. Weight values in last column of <table 2> were calculated with the equation (3). Links to the page 8 have the smallest weight value.

With this transition matrix, we can calculate each page's rank value. (Figure 2 and 3) show the change of the PageRank values and Hyperlink Information values in the each iteration respectively. We can find that the PageRank value and the Hyperlink Information values converge on certain values of both algorithms. <Table 3> shows the result value and each page's percentage. And last column of it shows the rate of difference between existing PageRank algorithm and Hypertext Information algorithm.

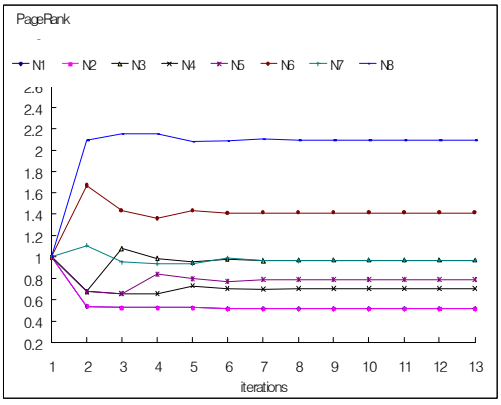

Fig. 2. Values in stable status in PageRank algorithm

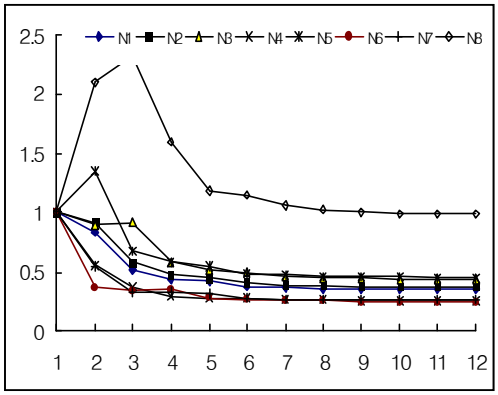

Fig. 3. Hyperlink Information values for the structure

In <Fig. 4> the PageRank (PR) and the Hyperlink Information (HI) and their corresponding ratios are represented. By comparison, we can see that the HI and the RatioHI give relatively higher values rather than PR in detecting the intentionally biased web page N8, even if the node has many inlinks. We enhanced the experimental result of [18] by using a Randomized weight onto the Hypertext Information.

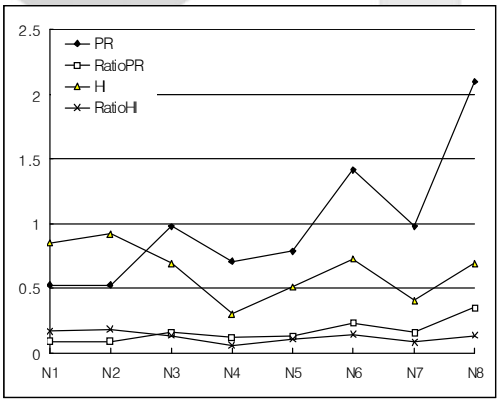

Fig. 4. Trajectories for PageRank and Hyperlink Information

V. Conclusions and future works

The web spam refers to hyperlinked pages on the web that are created with the intention of misleading search engines. The goal of this paper is how the PageRank algorithm and our approach work on these intentionally biased web pages and links and how efficiently these web pages can be detected. It is one of the most significant problems in the Web search engine that can result generate the best output to the user's submitted query and can effectively avoid the intentionally biased web pages. We discovered that the intentionally biased web page was exploiting the limitations of the PageRank based search engine's algorithm. In order to solve the problem originating from link based spamming, we modified the PageRank algorithm using the similarity between link contexts and hypertext information that can be generalized to the context based measure[17]. Using this proposed approach, the chance of intentionally biased web pages getting high ranks than deserved will be detected, and we can reinforce search accuracy.

▶Acknowledgements

This work was partially supported by the Korea Science and Engineering Foundation (KOSEF) through the Advanced Information Technology Research Center (AITrc).

REFERENCES

[1] Gyngyi, Z., Garcia-Molina, H., and Pedersen, J., "Combating Web Spam with TrustRank," In Proc. VLDB, pp. 576-587, 2004.

[2] Lee, W., "Semantic Access Path Generation in Web Information Management," Journal of The Korea Society of Computer and Information, Vol.8, No.2, pp.51-56, 2003.

[3] Halkida, M., Nguyen, B., Varlamis, I., Vazirgiannis, M., "THESUS: Organizing Web document collections based on link semantics," The VLDB Journal (12), pp. 320-332, 2003.

[4] Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A. and Rachavan, S., "Searching the Web," ACM Transactions on Internet Technology, Vol.1, No.1, pp.2-43, 2001.

[5] Miller, J., Rae, G. and Schaefer, F., "Modifications of Kleinberg's HITS Algorithm Using Matrix Exponentiation and Web Log Records," In Proc. ACM SIGIR, pp.444-445, 2001.

[6] Haveliwala, T., "Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search," IEEE TKDE, Vol. 15, No. 4, pp. 784-796, 2003.

[7] Brin, S., Page, L., "The Anatomy of a Large-Scale Hypertextual Web Search Engine," In Proc. WWW, pp. 107-117. 1998.

[8] Wang, Y., DeWitt, D., "Computing PageRank in a Distributed Internet Search Engine System," In Proc. VLDB, pp.420-431, 2004.

[9] Caldo, P., Ribeiro-Neto, B., Ziviani, N., "Local versus Global Link Information in the Web," ACM TOIS, Vol. 21, No. 1, pp. 4263, 2003.

[10] Phelps, T. and Wilensky, R., "Robust Hyperlinks: Cheap, Everywhere, Now," In Proc. DDEP/PODDP, pp. 28-43, 2000.

[11] Lu, W., Chien, L. and Lee, H., "Anchor Text Mining for Translation of Web Queries," In Proc. ICDM, pp. 401-408, 2001.

[12] Guo, D., Berry, M., Thompson, B., Bailin, S., "Knowledge-Enhanced Latent Semantic Indexing," Information Retrieval, Vol.6, No.2, pp.225-250, 2003.

[13] Castelli, V., Thomasian, A. and Li, C., "CSVD: Clustering and Singular Value De-composition for Approximate Similarity Search in High-Dimensi onal Spaces, IEEE TKDE, Vol. 15, No. 3, pp. 671-685, 2003.

[14] Ha, C., Youn, B., Hryu, K., "A Research on User's Query Processing in Search Engine for Ocean Using the Association Rules," Journal of The Korea Society of Computer and Information, Vol.8, No.2, pp.8-15, 2003.

[15] G. Nivasch, "Cycle detection using a stack," Information Processing Letters, Vol. 90, No. 3, pp. 135-140, 2004.

[16] L. Wookey, G. James, "Semantic Hierarchical Abstraction of Web Site Structures for Web Searchers," Journal of Research and Practice of Information Technology, Vol. 36, No.1, pp. 71-82, 2004.

[17] L. Wookey, K. Shin, S. -H. Kang, "Structuring Web with Semantic Hypertext Algorithm," In Proc. CITSA, Florida, USA, July 2004. pp. 257-262.

이 우 기

서울대학교 산업공학과 학사, 석사, 박사 취득

카네기멜런대학교 MSE 과정수료,

University of British Columbia 교환 교수,

ISI Canada TEFL deploma

2004년 한국경영과학회 최우수논문상 수상

현재 성결대학교 컴퓨터학부 부교수 재직

<관심분야> 정보검색, Web Structure Mining, 데이터웨어하우

저 자 소 개