A STUDY ON RANKING METHOD IN RETRIEVING WEB PAGES BASED ON CONTENT AND LINK ANALYSIS: COMBINATION OF FOURIER DOMAIN SCORING AND PAGERANK SCORING

Diana Purwitasari [1]

[1]Department of Informatics, Faculty of Information Science Technology
Institut Teknologi Sepuluh Nopember (ITS) Surabaya
Email: diana@its-sby.edu

Ranking module is an important component of search process which sorts through relevant pages. Since collection of Web pages has additional information inherent in the hyperlink structure of the Web, it can be represented as link score and then combined with the usual information retrieval techniques of content score. In this paper we report our studies about ranking score of Web pages combined from link analysis, PageRank Scoring, and content analysis, Fourier Domain Scoring. Our experiments use collection of Web pages relate to Statistic subject from Wikipedia with objectives to check correctness and performance evaluation of combination ranking method. Evaluation of PageRank Scoring show that the highest score does not always relate to Statistic. Since the links within Wikipedia articles exists so that users are always one click away from more information on any point that has a link attached, it it possible that unrelated topics to Statistic are most likely frequently mentioned in the collection. While the combination method show link score which is given proportional weight to content score of Web pages does effect the retrieval results.

Keywords: *ranking score, content analysis, link analysis, Fourier Domain scoring, PageRank scoring*

1 INTRODUCTION

With the rapid spread of the Internet, Web resources continue to grow and it has become increasingly difficult for users to find information on the Web. Under these circumstances, Web search engines help users to query useful information that satisfies their needs. Web search engine is an information retrieval, a search process, within linked document collection, the Web pages. The ranking module is an important component of the search process which sorts through too many or even until thousands relevant pages. This module takes relevant pages based on user queries and ranks them according to some defined rules. An ordered list as a result of ranking module filters relevant pages, makes relevant pages less and manageable for users. Distinguished from non-linked document collections, Web pages collection has additional information inherent in the hyperlink structure of the Web. For each relevant Web page, score of *link analysis* can be combined with score of *content analysis* to improve the quality of search results.

When link analysis, i.e. HITS [1] and PageRank [2], strikes to information retrieval fields, Web search have improved dramatically and nearly all major search engines now combine link analysis score with the usual information retrieval scores, including Google[1] and Ask.Com[2].

Initiated by works on link analysis, in fact it almost happened in the same time, some studies have suggested that combining link-based and content-based information can improve the retrieval performance of Web search [3] [4] [5]. The main reasons seem to be the users themselves are frequently uncertain about what exactly they are looking for such that they frequently provide short and imprecise queries. It makes the task of finding relevant information in Web search engines is hard. Therefore, the use of any source of additional available information beside relying on user queries must be considered in an attempt to improve the retrieval results.

Most of systems that do search process use variations of Boolean or vector space model to do ranking [6]. The concept behind vector space model is to convert each document and the query into vectors so they can be easily compared. The problem is any spatial information contained within is lost. Once converted into vectors, the number of times each term appears is represented, but the positions are ignored. In Fourier Domain Scoring, FDS, [7] rather than storing only frequency of terms, it stores signal of terms that show how terms are spread out.

Motivated by aforementioned reasons, in this paper we report our studies about ranking score of Web pages combined from *link analysis*, PageRank Scoring, and *content analysis*, Fourier Domain Scoring to improve the quality of search results. We have studied and then implemented the combination ranking method using Java programming. We do experiments to check correctness and performance evaluation of the implementation.

The structure of this report is written from our comprehension view about ranking method. Next sections are described about subjects related to ranking method: Fourier Domain Scoring and PageRank Scrong. Afterwards we describe about our implementation, then we report our experiments and close with some conclusions.

2 FOURIER DOMAIN SCORING, FDS

A method entitled with Fourier Domain Scoring, FDS, have been proposed to give scores of relevance to documents when related to a specific query [7] which can retain the document spatial information. This is illustration of the concept in FDS. If user input a query of "distance learning system", the retrieved results will contain terms of "distance", "learning", "system", or their combinations. But

[1]use link analysis of PageRank, available in `http://www.google.com/`

[2]use similar link analysis to HITS, available in `http://www.ask.com/`

what user really needs is the one that has text phrase of "distance learning system". Therefore, spatial or position information must be used to make sure that the query terms appear together in a certain orders.

FDS tries to capture the location of terms in documents. While VSM gives a relevance score based on frequency of term occurrences in documents, FDS not only justifies the occurrences by observing magnitude values but also compares the positions of terms by looking into phase values. In FDS, occurences and positions of terms are called as magnitude and phase values of terms signals. Relevant document to a query should have large magnitude, and corresponding phase of terms equal to query terms should be similar.

There are some scenarios to calculate magnitude, phase and how to represent them as a content score of document [7]. We had studied about Fourier Domain Scoring in previous works using both standard data set in information retrieval reasearch fields [8] and real data set of Web pages [9]. Based on the studies, we will use number of bins = 8, Sum Magnitudes to calculate magnitude, Zero Phase Precision to calculate phase, and Sum All Components to produce content score for retrieving relevant documents in searching processes. Description of all steps to calculate content score here have been already adapted to our problem domain.

2.1 Calculate Term Weight

A document may contain at least hundred of terms which eventually results to term signals in almost the same amount. The terms should be grouped into bins to reduce the size of term vector. Let t be a certain index term in document d and B is number of bins. Based on our previous works, we set $B = 8$. A term weight $\omega_{d,t,b} > 0$ is associated to number of term occurrences for term t inside document d in location bin b, where $b \in \{0, \ldots, 7\}$. Let N be total number of documents in the collection and n_t be the number of documents in which index term t appears. Then weighting schemes of $\omega_{d,t,b}$ which derived from TFIDF schemes [6] will be:

$$\omega_{d,t,b} = \frac{freq_{d,t,b}}{\max\limits_{i,j} freq_{d,t=i,b=j}} \cdot \log \frac{N}{n_t} \qquad (1)$$

where $freq_{d,t,b}$ be number of term occurrences for term t within document d location bin b.
While $\max_{i,j} freq_{d,t=i,b=j}$ is a maximum value of number of term occurrences for certain term i in certain location bin j from all terms inside document d. The division with maximum value will normalize $\omega_{d,t,b}$ into a unit value, $0.0 \leq \omega_{d,t,b} \leq 1.0$.

2.2 Transform into Fourier Domain

After calculation of $\omega_{d,t,b}$ and makes a term signal for each term, $[\omega_{d,t,0} \ldots \omega_{d,t,7}]^T$, then transform those values of time or spatial domain into frequency domain. Fourier transform defines a relationship of signals in time or spatial domain with its representation in frequency domain known as a (Fourier) spectrum. A spectrum is made up of a number of frequency component with real and imaginary parts for each frequency component. A Fourier transform in FDS would produce the following mapping [7]:

$$\{\omega_{d,t,b}\} \rightarrow \{v_{d,t,b}\} = \{H_{d,t,b} \exp(i\phi_{d,t,b})\} \qquad (2)$$

Note, b symbol in left and right equation do not show the same notation as number of bins. The left part shows index of *bin* but the right part shows index of *frequency component*. To analyze signals in frequency domain, real and imaginary parts can be converted into magnitude, $H_{d,t,b}$, and phase, $\phi_{d,t,b}$, of each frequency component. Here i is $\sqrt{-1}$ to show imaginary part. $v_{d,t,b}$ is a projection of $\omega_{d,t,b}$ in term signal $[\omega_{d,t,0} \ldots \omega_{d,t,7}]^T$ onto a sinusoidal wave of frequency component b.

Discrete Fourier Transform to do a projection of $\omega_{d,t,b}$ onto $v_{d,t,b}$ is shown as below [7]: (Note here to easily differentiate, frequency component b is changed into β so that the projection result will be $v_{d,t,\beta}$)

$$v_{d,t,\beta} = \sum_{b=0}^{7} \omega_{d,t,b} \exp^{-\frac{2\pi i}{8}\beta b}, \qquad \beta \in \{0, \ldots, 7\} \qquad (3)$$

The spectral component number β is an element of the set $\{0, \ldots, 7\}$. The upper bound of bin and frequency component follow the results from our previous works with number of bin $B = 8$ so that number of frequency component also $\beta = 8$.

Example 1 *Let a term signal of* $[\omega_{d,t,0} \ldots \omega_{d,t,7}]^T$ *be* $[1\,1\,0\,0\,0\,2\,0\,0]^T$. *Calculate magnitude and phase for frequency component* $\beta = 1$, $v_{d,t,1}$.
First is to calculate the real part, $re_{d,t,1}$.

$$\begin{aligned} re_{d,t,1} &= \sum_{b=0}^{7} \omega_{d,t,b} \cos(-\frac{\pi}{4}b) = \sum_{b=0}^{7} \omega_{d,t,b} \cos(45°b) \\ &= \cos 0° + \cos 45° + 2\cos 225° = 0.2929 \end{aligned}$$

Second is to calculate the imaginary part, $im_{d,t,1}$.

$$\begin{aligned} im_{d,t,1} &= \sum_{b=0}^{7} \omega_{d,t,b} \sin(-\frac{\pi}{4}b) = \sum_{b=0}^{7} \omega_{d,t,b} \sin(-45°b) \\ &= \sin 0° - \sin 45° + 2\sin 225° = 0.7071 \end{aligned}$$

Then, calculate its magnitude, $H_{d,t,1}$.

$$H_{d,t,1} = (re_{d,t,1}^2 + im_{d,t,1}^2)^{\frac{1}{2}} = 0.7654$$

Finally, calculate its phase, $\phi_{d,t,1}$. *For* $\beta = 1$, *since* $re_{d,t,1} > 0, im_{d,t,1} > 0$ *thus* $\phi_{d,t,1}$ *should be in a range of* $0 < \phi_{d,t,1} < \frac{\pi}{2}$ *or* $0 < \phi_{d,t,1} < 1.5708$.

$$\phi_{d,t,1} = \tan^{-1} \frac{im_{d,t,1}}{re_{d,t,1}} = 1.1781$$

2.3 Calculate Magnitude

Sum Magnitudes $H_{d,b}^{m}$ only takes account of magnitude, $H_{d,t,b}$, to ensure more weight is given to document d with

more occurrences of query terms. Let $\omega_{q,t}$ be a term weight in query q calculated with:

$$\omega_{q,t} = \frac{freq_{q,t}}{\max freq_{q,t}} \cdot \log \frac{N}{n_t} \quad (4)$$

where $freq_{q,t}$ be number of term occurrences for term t within query q. While $\max freq_{q,t}$ is a maximum value of number of term occurrences for certain term from all terms inside the query q. To calculate $H^m_{d,b}$, magnitude of all terms within document d that equal to query terms are multiplied by their correlational query term weight $\omega_{q,t}$.

$$H^m_{d,b} = \sum_{t \in T} H_{d,t,b} \cdot \omega_{q,t} \quad (5)$$

where T is a set of query terms.

2.4 Calculate Phase

Zero Phase Precision $\Phi^z_{d,b}$ only includes phase, $\phi_{d,t,b}$, filtered by nonzero magnitude value. Because term with zero magnitude value means that the term does not exist and its phase value could be left out. To calculate $\Phi^z_{d,b}$, phase of all filtered terms within document d that equal to query terms will be summed and averaged with total number of query terms, $\#(T)$.

$$\Phi^z_{d,b} = \sqrt{\left(\frac{\sum_{t \in T} \cos \phi_{d,t,b}}{\#(T)}\right)^2 + \left(\frac{\sum_{t \in T} \sin \phi_{d,t,b}}{\#(T)}\right)^2} \quad (6)$$

2.5 Calculate Content Score

After magnitude, $H^m_{d,b}$, and phase, $\Phi^z_{d,b}$, of frequency components have been obtained, the next step is to combine them to create score vector for each frequency component by multiplication.

$$s_{d,b} = H^m_{d,b} \cdot \Phi^z_{d,b} \quad (7)$$

To get content score of document, FDS_d, score of each frequency component will be summarized. However since Nyquist-Shannon sampling theorem states that highest frequency component found in a real signal is equal to half of the sampling rate, it implies that to analyze term signal we would only need to examine half of frequency components.

$$FDS_d = \sum_{b=1}^{4} s_{d,b} \quad (8)$$

Here, the zero-frequency components (representing the mean intensity of the signal) as well as the (redundant) negative ones are omitted.

3 PAGERANK SCORING, PRS

To improve quality of search results, we use PageRank [2] as a technique that analyze additional information inherent in the hyperlink structure of Web pages to determine the link score, PRS_d. Then, content score is combined with link score to determine score of documents in the collection. PageRank measures relative importance of each hyperlinked document within the collection and gives a weight score to that importance. It means that a document which is *linked-to* by many documents receives a high rank. Here term of document and term of Web page refer to the same meaning, data within hyperlinked collection.

A Web page will have a higher score if it receives more recommendations from other pages (read: has *in*-links). Since status of the recommender is also important, therefore weight of each inlink should be tempered by total number of recommendations made (read: has *out*-links).

PageRank of a page P_i, denoted $r(P_i)$, is sum of PageRanks of all pages pointing into P_i. Let B_{P_i} is set of pages pointing into P_i and $|P_j|$ is number of outlinks from page P_j. The PageRank of inlinking pages $r(P_j)$ is tempered by number of recommendations made by P_j, denoted $|P_j|$. Since in the beginning the $r(P_j)$ values, the PageRanks of pages inlinking to page P_i, are unknown, it is assumed that initially all pages have equal PageRank. Using iterative procedure, let $r_{k+1}(P_i)$ be PageRank of page P_i at iteration $k+1$, then

$$r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{|P_j|} \quad (9)$$

This process is initiated with $r_0(P_i) = \frac{1}{n}$ for all pages P_i where n is the number of pages indexed from the Web. Then iteratively calculate PageRank scores until converge to some final stable values. We use Least Square Error[10] where until differentiation of PageRank score at iteration $k+1$ with its previous value is smaller than threshold value ϵ, the iterative calculation of PageRank should continue.

$$\sum_{i=1}^{n} [r_{k+1}(P_i) - r_k(P_i)]^2 \leq \epsilon \quad (10)$$

In this paper, threshold value is decided as $\epsilon = 1 \times 10^{-8}$.

In PageRank model, hyperlink structure of the Web is defined as a massive directed graph with nodes and directed edges represent Web pages and hyperlinks. Then the Web directed graph defines *adjacency matrix* $H_{(n \times n)}$ having $h_{ij} = \frac{1}{|P_i|}$ if there is a link from page P_i to page P_j, and 0, otherwise.

Rather than computing PageRank of one page at a time, use matrices to compute a PageRank row vector $\pi^T_{(1 \times n)}$ which holds PageRank values for all pages in the index. Let $\pi^{(k)T}$ is PageRank vector at k^{th} iteration and (9) can be written compactly as

$$\pi^{(k+1)T} = \pi^{(k)T} H \quad (11)$$

with $\pi^{(0)T} = \frac{1}{n} e^T$ where e^T is the row vector of all 1s.

3.1 Markov Model of the Web

The mathematical component of PageRank vector is resembled to the power method applied to a Markov chain

with *transition probability matrix* H[10]. A unique positive PageRank vector exists when matrix H is **stochastic**, **irreducible** and **primitive**. Brin and Page force some adjustments to *adjacency matrix* $H_{(n\times n)}$ since, unfortunately, matrix representation of real Web graph does not always have those sufficient conditions.

Definition 1 *A stochastic matrix $H_{n\times n}$ is a non negative matrix in which sum of each row is equal to 1. Matrix H is said to be a non negative matrix whenever each element $h_{ij} \geq 0$.*

Matrix H is always a non negative but might be not stochastic because there are pages with no outlinks, $h_{ij} = 0$, and will make sum of row i is not equal to 1, $\sum_{j=1}^{n} h_{ij} = 0$. Based on assumption that after reaching no outlinks pages, user can go to any other page at random, the 0^T rows are replaced with $\frac{1}{n}e^T$ in matrix H. Stochastic adjustment makes (11) into

$$\pi^{(k+1)T} = \pi^{(k)T}\left[H + a\frac{1}{n}e^T\right] = \pi^{(k)T}S \qquad (12)$$

Binary column vector $a_{n\times 1}$ is called dangling node vector where $a_i = 1$ if page i has no outlinks and 0, otherwise.

Figure 1(a) shows a sample of small Web graph where nodes and edges represent Web pages and links between pages, i.e. node 1 has 2 outlinks and 1 inlink. Since page P_2 has no outlinks, stochastic adjustment is needed and the result is shown in Fig. 1(b). A non negative matrix H is changed into a stochastic matrix S.

Example 2 *Matrix H from Figure 1(a) is shown as below:*

$$H = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

$$S = H + \frac{1}{6}\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Here, to change into stochastic adjustment, the 0^T row or the second row are replaced with $\frac{1}{6}e^T$.

Definition 2 *Matrix $H_{n\times n}$ is irreducible if and only if for each pair of indices (i, j) there is a sequence path from every other pairs of indices, i.e. path from page i to page j exists in matrix H such that $h_{ik_1}h_{k_1k_2}\cdots h_{k_tj} \neq 0$.*

After following the hyperlinks in Fig. 1(b) and finally page P_4, P_5 or P_6 has been reached, user would never be able to return to P_1, P_2 or P_3. User is trapped into bouncing endlessly between cycle of P_4, P_5 and P_6. Figure 1(b) shows a reducible graph because there is no sequence paths from P_4, P_5 and P_6 to P_1, P_2 or P_3.

Based on assumption that user might get bored to follow the hyperlink structure of the Web and randomly jump to another Web page, irreducible adjustment is done. Brin and Page[2] invented Google matrix G to accomodate that situation. Irreducible adjustment changes (12) into

$$\pi^{(k+1)T} = \pi^{(k)T}\left[\alpha S + (1-\alpha)\frac{1}{n}ee^T\right] = \pi^{(k)T}G \qquad (13)$$

Parameter α controls proportion of time user follows hyperlinks while $(1-\alpha)$ is proportion of time user jumps to another page randomly. Multiplication column vector e and row vector e^T with a constant value $\frac{1}{n}$ will make a transition probability matrix where there is always a sequence path for each pair of indices. That guaranteed-irreducible-transition-probability matrix combines with stochastic matrix S will result into a stochastic-irreducible matrix G.

Definition 3 *Matrix $H_{n\times n}$ is defined to be primitive when H is a non negative irreducible matrix and $H^m > 0$ for some $m > 0$.*

Because all elements $g_{ij} > 0$ in transition probability matrix G, then just for $m = 1$ will make $G^m > 0$. Therefore irreducible adjustment also guarantees that matrix G is sufficient to conditions of stochastic, irreducible and primitive.

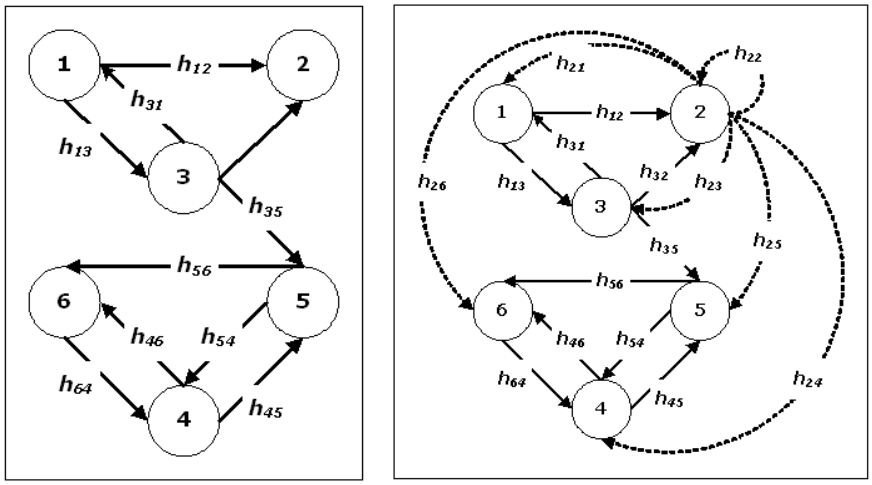
Example 3 *Use Figure 1(a) as a sample of small Web graph. Matrix G with $\alpha = 0.9$ will be:*

$$G = 0.9H + \left(0.9\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + 0.1\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}\right) \times \frac{1}{6}\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

After first iteration, the result of PageRank score PRS_d will be $\pi^{(1)T} =$ $[0.037 \quad 0.054 \quad 0.042 \quad 0.375 \quad 0.206 \quad 0.286]$. *It means that most likely user will visit P_4 because its link score is the highest, $PRS_{P_4} = 0.375$.*

4 IMPLEMENTATION

We have implemented a ranking method with Fourier Domain Scoring and PageRank Scoring. FDS will make sure that terms occurring close together in a Web page are weighted higher than terms occurring far apart. For a multi-keywords searching, the situation is more complicated than a single keyword query. Keywords occurring close together should be weighted higher, moreover to calculate the distances between them would require a lot of calculations at query time. PRS is query-independent since its calculation could be done eventhough no queries entered from user. FDS needs magnitude and phase information of each keyword, but only magnitude calculation is the one that query-dependent which initiates in real-time when a user enters a query. On the other hand, precalculated phase is



(a) Original Web graph

(b) After stochastic adjustment

Figure 1: Adjustment on sample real Web graph to have sufficient conditions for Markov model

query-independent. Therefore FDS could easily be implemented in ranking system which using PRS, where the system will previously calculate phase along with link score of Web pages.

We design our ranking function so that content score in each page is proportioned with its link score. After normalization for each score of web pages, content score is combined with link score by multiplication to give an overall score S_d for each relevant page for a given query.

$$S_d = FDS_d \times PRS_d \tag{14}$$

5 EXPERIMENTS

Our experiments in this report use collection of Web pages from Wikipedia[3]. We have two objectives, correctness and performance evaluation. We had checked the correctness of Fourier Domain Scoring in our previous works [9] [8]. For the correctness of PageRank Scoring, we test PRS implementation with Harvard data set[4] contains representation of 500 Web pages from Harvard site[5]. PageRank scores of Web pages are illustrated in Fig. 2 which shows the first node has the highest score value. That node is the main page of Harvard site. It shows that most of Web pages in Harvard site have reference link to the main page as usually happened in any existing site in the Web which always gives opportunity for users to go back to the home of site.

Began with preprocessing using Oracle Text[11] to extract terms and then creating an inverted index, our ranking system–FDS module[9], PRS module and FDS×PRS module–is developed using Java. Data collection, d_1, is retrieved by crawling from main page *Statistic*[6] using link depth level = 1 and other data collections, d_2, d_3, d_4, are using same main page but link depth level = 2.

5.1 Analyze PRS

We have experiments for calculating PageRank scores with four selected data set collections, $d_{i\in 1\cdots 4}$, $n_{d_1} = 34$, $n_{d_2} = 100$, $n_{d_3} = 500$, $n_{d_4} = 1083$). We use third party crawler engine and save copies of Wikipedia articles in local disk. All our experiments are executed as offline processes. Each calculation always assumed that initial PageRank score of all Web pages in the index is equal and $\alpha = 0.85$[10].

Figure 3 shows PageRank score of web pages from *node id*$\in\{1..32\}$ in data collection d_1 because there are Web pages refer to the same URLs. This is happened because Wikipedia often makes redirection links from old articles to better coverage articles. After redirection, URLs of old articles are not removed so that in crawling process those URLs are still being parsed. Convergence after 33 iterations, Web page with the highest PageRank score is titled with *Statistics* [7].

Figure 4 shows PageRank score from data collection d_1, added with some other linked pages and named as d_2 with actual node number is 94. In data collection d_2, *Mathematics*[8] Web page has the highest PageRank score, followed with subject *Algorithm* and *Statistics* after 25 iterations. A little bit different result from our expectation on Statistic data collection, but after all statistics is related to mathematical science and that makes subject *Mathematics* is also important.

Figure 5 shows PageRank score and converges after 12 iterations from data collection d_3 with actual node number is 466. Web page with the highest PageRank score has title *2006*[9]. At this point, the highest PageRank score does not

[3] http://en.wikipedia.org/

[4] available in http://www.mathworks.com/moler/chapters.html

[5] http://www.harvard.edu

[6] available in http://en.wikipedia.org/wiki/statistic

[7] http://en.wikipedia.org/wiki/Statistics

[8] http://en.wikipedia.org/wiki/Mathematics

[9] http://en.wikipedia.org/wiki/2006

related at all with topic Statistic. The problem is caused by characteristic of Wikipedia site where anyone can create and modify content of web pages, such as make any keyword becoming hyperlink to another page. The links within Wikipedia articles exists so that users do not need to cover common ground in depth when learning subjects on current article. Instead, the users are always one click away from more information on any point that has a link attached. In this case, subject around year of *2006* is most likely frequently mentioned and important within data collection d_3. The next highest PageRank score belongs to Web page with title *English language* and *Mathematics*.

Figure 6 shows PageRank score after 12 iterations from data collection d_4 with actual node number is 993. Web pages with highest PageRank score are, in decreasing order, subjects of *Mathematics*, *2006*, *Wikipedia* and *Statistics*.

5.2 Analyze FDS×PRS

Figure 7 shows Web page scores from data collection d_1 that relevant to query "probability distribution" with descending ordered. FDS×PRS scores are generally propotional to PRS scores. Web page with not to good content score but higher link score (*node id = 32*) has better final score than the one with higher content score but lower link score (*node id = 31*). However eventhough a Web page has higher link score but its content score is too low (*node id = 3*), then its final score becomes lower too.

In experiment to compare our ranking system with existing search engine results, we define query "probability distribution" with some conditions:
(a) search from data collection using combination ranking method of FDS×PRS. (b) search from Google with limitations to retrieve English web pages and only from Wikipedia domain. (c) search from Wikipedia search feature.
Comparison is made by observing relevant pages with precisions achieved after retrieving 60 Web pages (or the percentage of relevant and retrieved items after inspecting the first 60 Web pages).

FDS×PRS's results give precision 0.47 compared to Google's. Experiments show that Google gives better precision since it does not only record occurrence and similarity of multiple query keywords in Web pages for content scores, but also takes into account things such as whether the query keywords appear in the title or deep in the body of a Web page[2]. Google considers more weight on query keywords that appear in title of Web pages while FDS×PRS does not. Google might also consider query keywords position in title part because Web pages with "probability" in title have higher ranks than Web pages with "distribution".

Better precision, 0.50, is resulted from comparison between FDS×PRS and Wikipedia. In Statistics theory there are kinds of distribution like Binomial, Normal, etc. Some of Wikipedia Web pages have links to other pages that explain those distributions, but they are only written as "Binomial" or "Normal" not "Binomial Distribution" or "Normal Distribution" inside contents of Web pages. FDS×PRS gives higher ranks to those Web pages eventhough only a few keywords of "probability distribution" exists.

6 CONCLUSIONS AND FUTURE WORKS

We have experimented with Fourier Domain Scoring and PageRank Scoring to check the correctness and the performance evaluation of ranking methods. Correctness evaluation of PageRank Scoring show that the highest score does not always relate to Statistic. Characteristic of Wikipedia site is anyone can create and modify content of web pages, such as make any keyword becoming hyperlink to another page. The links within Wikipedia articles are created so that users do not need to cover common ground in depth when learning subjects on current article and can always be one click away from more information on any point that has a link attached. It it possible unrelated topics to Statistic are most likely frequently mentioned in the collection, i.e. Web page entitled with certain year as a page contains lists of articles (note this kind of page has sole purpose to organize Wikipedia articles by subject or alphabetically to help users to make the best use of Wikipedia through searching and traversing in a loose hierarchy for further related information).

Performance evaluation of our ranking system which using scoring method based on combination content and link analysis of Web pages show improvement on the retrieval results. However since there are many rules that could be used to give content score to each relevant page, our ranking system needs to take into account on more things such as the appearance of query keywords in a page. For example, in Wikipedia Web pages, keywords that appear in table-of-contents or reference division should be weighted more.

References

[1] Kleinberg, J.M.: *Authoritative Sources in a Hyperlinked Environment*. Journal of the ACM **46**(5) (1999) 604–632

[2] Page, L., Brin, S., Motwani, R., Winograd, T.: *The PageRank Citation Ranking: Bringing Order to the Web*. Technical report, Stanford Digital Library Technologies Project (1999)

[3] Bharat, K., Henzinger, M.R.: *Improved Algorithms for Topic Distillation in a Hyperlinked Environment*. In: SIGIR '98: Proc. of the 21st Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, Melbourne, Australia (1998) 104–111

[4] Silva, I., Ribeiro-Neto, B., Calado, P., Moura, E., Ziviani, N.: *Link-based and Content-based Evidential Information in a Belief Network Model*. In: SIGIR '00: Proc. of the 23rd Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, Athens, Greece (2000) 96–103

[5] Jin, R., Dumais, S.: *Probabilistic Combination of Content and Links*. In: SIGIR '01: Proc. of the 24th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, New Orleans, Louisiana, United States (2001) 402–403

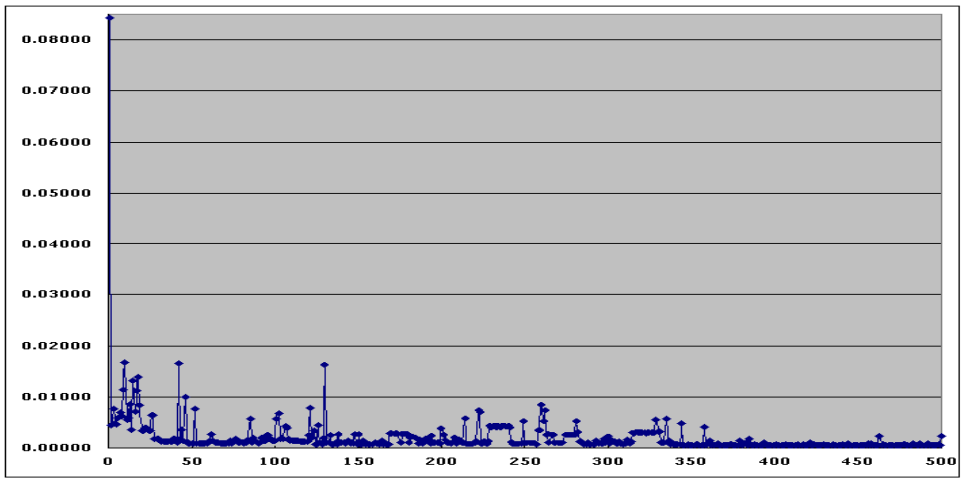


Figure 2: PageRank score for data collection of Harvard Web pages; $n = 500$ documents

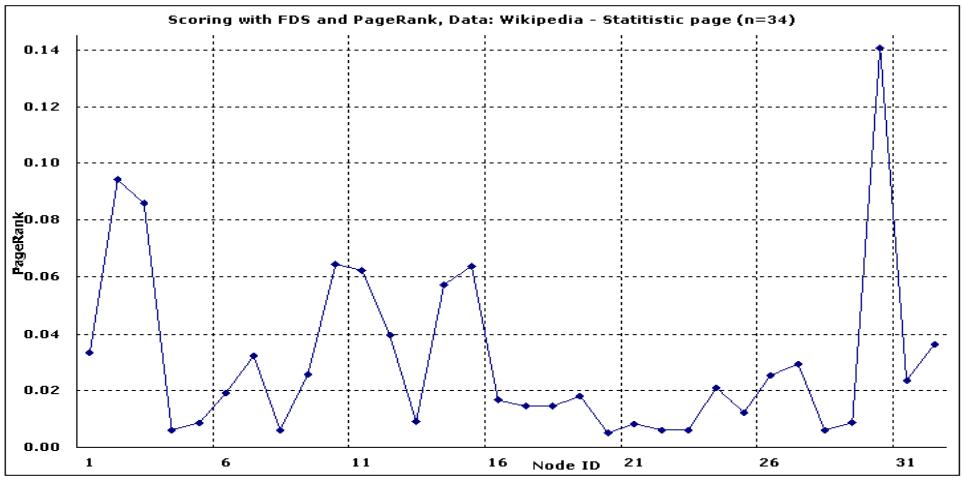


Figure 3: PageRank score for data collection of Wikipedia Web pages, d_1; $n = 34$ documents

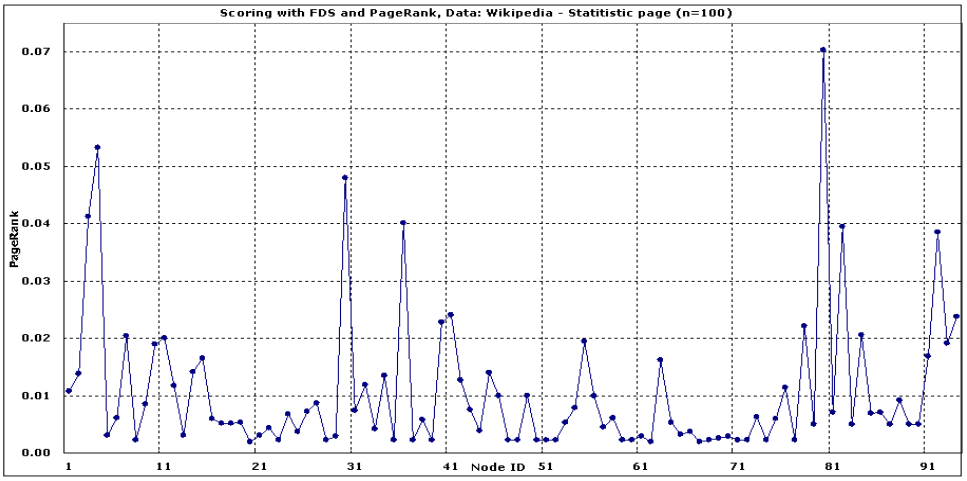


Figure 4: PageRank score for data collection of Wikipedia Web pages, d_2; $n = 100$ documents

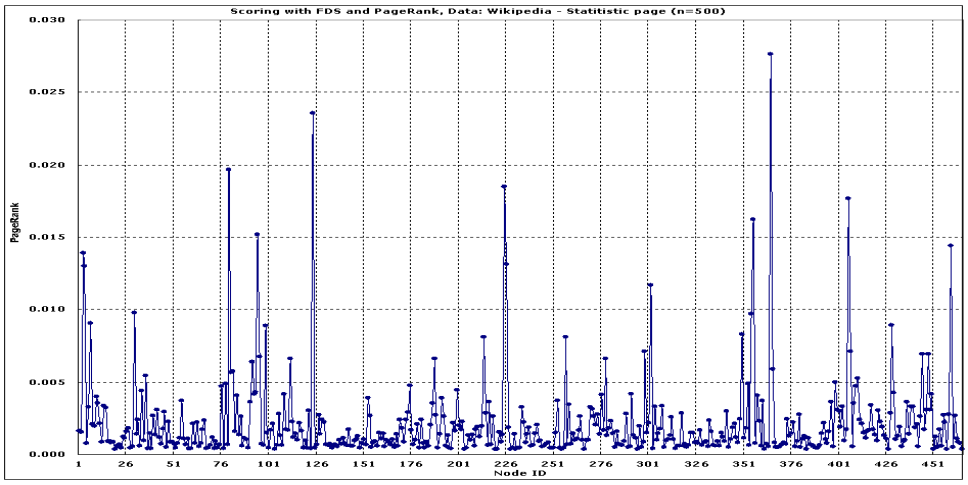


Figure 5: PageRank score for data collection of Wikipedia Web pages, d_3; $n = 500$ documents

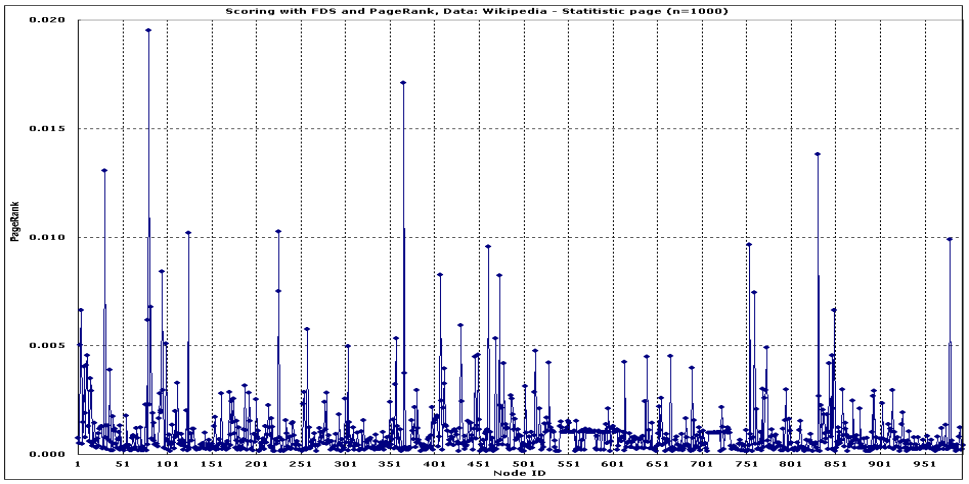


Figure 6: PageRank score for data collection of Wikipedia Web pages, d_4; $n = 1083$ documents

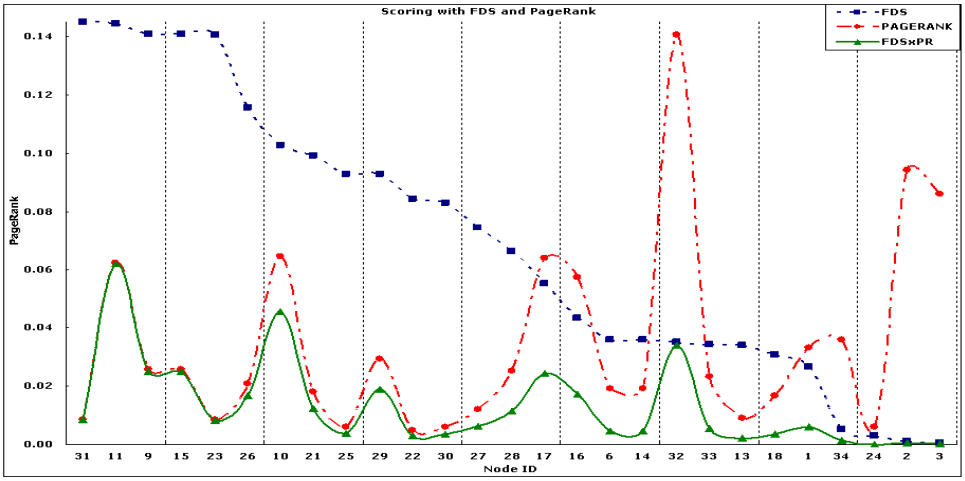


Figure 7: Fourier Domain Scoring and PageRank Scoring effect score of Web pages

[6] Yates, R.B., Ribeiro-Neto, B.: *Modern Information Retrieval*. Addison-Wesley (1999)

[7] Park, L.A., Ramamohanarao, K., Palaniswami, M.: *Fourier Domain Scoring: A Novel Document Ranking Method*. IEEE Trans. on Knowledge and Data Engineering **16**(5) (2004) 529–539

[8] Purwitasari, D., Suciati, N., Soelaiman, R., Farida, D.: *Fourier Domain Scoring for Ranking Method in Small Data Set with Preprocessing Using Oracle Text*. In: ICTS'07: The 3rd Intl. Conf. on Information & Communication Technology and Systems, Surabaya, Indonesia (2007)

[9] Purwitasari, D., Okazaki, Y., Watanabe, K.: *A Study on Web Resources' Navigation for e-Learning: Usage of Fourier Domain Scoring on Web Pages Ranking Method*. In: ICICIC'07: Second Intl. Conf. on Innovative Computing Information and Control, Kumamoto, Japan, IEEE Computer Society (2007)

[10] Langville, A.N., Meyer, C.D.: *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press (2006)

[11] Oracle Technology Network: *Oracle Text in Oracle Database 10g release 2* (2005) `http://www.oracle.com/technology/products/text/index.html`.

[HALAMAN INI SENGAJA DIKOSONGKAN]