Diversity-Based HITS: Web Page Ranking by Referrer and Referral Diversity

Yoshiyuki Shoji and Katsumi Tanaka

Department of Social Informatics
Graduate School of Informatics, Kyoto University
Kyoto, Japan
{shoji, tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract. We propose a Web ranking method that considers the diversity of linked pages and linking pages. Typical link analysis algorithms such as HITS and PageRank calculate scores by the number of linking pages. However, even if the number of links is the same, there is a big difference between documents linked by pages with similar content and those linked by pages with very different content. We propose two types of link diversity, referral diversity (diversity of pages linked by the page) and referrer diversity (diversity of pages linking to the page), and use the resulting diversity scores to expand the basic HITS algorithm. The results of repeated experiments showed that the diversity-based method is more useful than the original HITS algorithm for finding useful information on the Web.

1 Introduction

As the World Wide Web continues to rapidly expand, both Internet users and the purposes of Web documents are becoming more extensive and diverse. Users of the Web are not only computer specialists but also ordinary people, and the content on the Web has accordingly become broader, including not just informative documents but also many sub-products of communication, personal diaries, and so on. This has made Web usage increasingly more complex.

Many major Web search engines use link analysis algorithms such as HITS and PageRank to rank Web documents. For example, Google uses PageRank, which calculates popularity scores. These methods focus on the number of linking documents of each page. The popularity score is determined by recursive calculation using the simple hypothesis that pages linked by many popular pages are popular pages. The HITS algorithm uses the number of linking and linked pages to calculate the Hubs and the Authority. However, these methods focus only on the number of linking pages and are therefore unsuitable for coping with the demands of complex usages of the Web.

For example, when a novice user wants to know what a "compiler" is and inputs the query "compiler" to a search engine, the result contains many different kinds of pages, such as dictionary pages (e.g., thesaurus), encyclopedia

pages (e.g., Wikipedia), introductory articles, commercial sites about specialized compilers, academic articles, and so on. All of them are popular documents and contain the term "compiler." However, the most popular documents are not necessarily useful for all users. In this case, dictionary, encyclopedia pages and introductory articles would be useful for novice and general users, but other pages would be more useful for a limited number of specialists. The problem here is that the existing link analytic methods score every document based on how many documents link to them instead of checking how it was linked. The number of linking documents expresses how popular the document is, but does not express why it is popular.

In this paper, we propose a new link analysis algorithm that considers not only the number of linking pages but also the diversity of linking pages and linked pages in order to consider the reason why the page is popular.

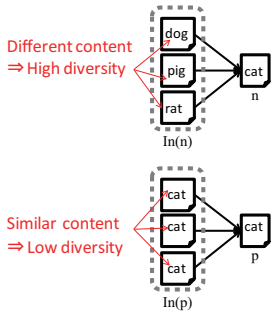

Fig. 1. Referrer diversity.

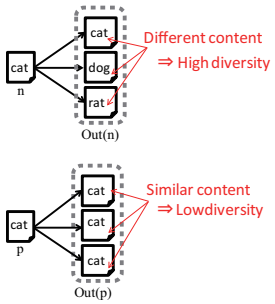

Fig. 2. Referral diversity.

Figure 1 shows an example of different ways a page is linked. There are two pages about cats, both of which are linked by three pages. Page n is the first page about cats, which is referred by documents on three different topics: a page about dogs, a page about pigs, and a page about rats. Page p is also linked by three pages, but all of them are about cats. Since each linking page has the same popularity score, PageRank gives the same popularity score to page p and page n because both of them are linked by the same number of pages. However, we suspect there is a big difference in the reason behind their popularity. Page n is linked by diverse pages. The authors of pages that link other pages may just be readers of the latter, who create the link after viewing such pages and finding them interesting. When the topic in the document reflects the interest of the author, a page linked by diverse pages must have a wide readership. An article with a diverse readership is assumed to contain information that can be interesting for many people: general information, universal information, and so on. Such information is useful not only for specialists but also for novice users. In contrast, page p has a biased, non-diverse readership, making it suitable for certain specialists or members of a specific community. The same can be said

of linking sites and navigating sites(see Fig. 2). In this case, the former linking page may have been made by a user who has a broad outlook and the latter by one whose range of interest is narrow. These examples demonstrate that by considering how the linking pages and linked pages are diverse, the search algorithm can create a ranking that depends on the reason behind the popularity.

We define diversity as the dispersion of a set of pages. When each page in the set has a different topic, the diversity score is high, and when all pages are similar, the diversity score is low. We expand the HITS algorithm, which is the standard existing link analysis algorithm, with two types of diversity score: referrer diversity and referral diversity. Referrer diversity means how pages that link the page are diverse and referral diversity means how pages linked by the page are diverse. The conventional HITS algorithm calculates Hubs, which means how the page links many good authorities, and Authorities, which means how the page is linked by many good hubs. We expand the concept of Hubs and Authorities with diversity by two simple hypotheses:

- The page linking diverse Authorities is a valuable Hub
  (This Hub can be created by a well-informed generalist who has a wide range of interests.)
- The page linked by diverse Hubs is a valuable Authority
  (It must be useful not only in a specific field.)

The rest of this paper is organized as follows. Section 2 describes related work. In Section 3, we introduce the diversity-based link analysis algorithm we developed. Section 4 describes our experiments, and Section 5 evaluates our method in light of the experimental results. We conclude this paper in Section 6.

2 Related Work

In this paper, we propose a link analysis ranking algorithm that considers the diversity of linking pages and linked pages to isolate general information. We adduce three related previous studies. Section 2.1 describes research on diversity, section 2.2 describes other link analytic ranking algorithms, and section 2.3 describes methods to identify general information.

2.1 Diversity

Diversity in information science is a very active research topic. Collective intelligence is discussed particularly actively since the interactive usage of Web sites has become more common. This is called Web 2.0. Surowiecki [1] presented conditions for data and methods to realize the wisdom of crowds in his 2005 book. In his view, data need to meet three requirements: “Diversity of opinion,” “Independence” and “Decentralization.” When perfect data are available, they should be handled with “Aggregation,” which means an algorithm to make up congregate data to knowledge. Link analytic ranking algorithms such as PageRank can

be assumed as a voting model of crowds that treat the link as an acceptance. Thus, any discussion on diversity should consider the validity of the ranking.

Diversity has also been discussed in the field on Information Retrieval. One of the most active research topics in this area is the diversification of Web search result pages. These studies tackle the problem of how many diverse pages can appear in the first result page [2] [3] [4]. The diversification of Web search results can be classified by two features [5]:

- Ambiguous query terms
- Available information sources

For instance, the query "jaguar" is an ambiguous query in that it can have several meanings, such as the car manufacturer, the animal, a personal name, and so on. To diversify the search result of "jaguar," ranking algorithms should rank the pages relevant to the different meaning of "jaguar," alphabetically and without overlaps. If the aim of the query is identified, it is a problem when the ranking contains the same kind of information. Search algorithms should therefore create rankings that cover different types of content.

The research areas that discuss diversity are not limited to information sciences but include sociology, ecology, life science, economics, and so on. Stirling [6] put forth three key factors on categorical diversity: "Variety," "Balance" and "Disparity." To come up with the optimal design for the calculation of diversity, we need to place emphasis on these three factors.

2.2 Link Analysis Algorithms

Here, we present the new link analysis-based algorithm we developed. There are many past examples that expand PageRank and HITS for each aspect.

For instance, the topic sensitive PageRank [7] deals with the topic of the query and documents. The TrustRank [8] algorithm uses PageRank for spam filtering on the basis of the simple theory that spam pages link to both good and spam pages but good pages link only to good pages. Another idea is building the concept of time and space into PageRank to measure the historic impact [9].

Alternative approaches featuring Web graph analysis have been proposed. The HITS algorithm [10] uses the link structure on the Web to locate communities in the Web graph. It models the Web graph as a bipartite graph and calculates the importance of Web pages by convergence calculation. It uses two types of scores: a Hub score and an Authority score. Authorities are pages that show good information, and they obtain a higher score if the page links many highly scored Hub pages. Hubs are pages that link to good Authority sites. A typical example of a good Hub site is a linking site or a good search result page. The generalized co-HITS algorithm [11] is a HITS-based algorithm that extends the conventional HITS algorithm from Web links to general bipartite graphs such as a paper-and-author pair. SALSA [12] has similar algorithms and makes bipartite graphs based on Hubs and Authorities and has a score calculated based on random walk. Fast random walk with restart [13] is similar, too, and is used to compute the similarity between nodes.

Our algorithm is a HITS-based algorithm with weighting. The unique part is that it defines the weight of the nodes depending on the diversity.

2.3 Finding Information for Novice Users

The aim of this research is to enable novice users to find useful information on the Web. There are a few previous studies that focus on the comprehensibility or specialty of a given document. Our method finds the information for beginners by means of a link analytic approach. There has also been much research on estimating the specialty of a document by a content analysis approach. In particular, estimating the specialty of terms included in documents is a hot topic.

In the field of natural language processing, several methods that extract the special terms (i.e., technical terminology and jargon) from documents have been proposed. Nakatani et al. [14] proposed a link analytic method using the Wikipedia category structure to extract special terms. These studies used big corpora or document sets of limited specialized fields and structural information to measure the specialty of a term. Our method does not aim to find special terms but rather special Web pages without using particular datasets. The existing methods can be used to increase the accuracy of our method in a complementary style. There is a previous similar study that aims to find comprehensible Web pages by the link analytic approach. Akamatsu et al. [15] proposed a TrustRank-based method built on one simple rule: comprehensible pages are more likely to link comprehensible pages. General pages that we want to find are similar to comprehensible pages, but this method is based on PageRank, which focuses only on the number of links and not on how they link.

3 Proposed Method

In this section, we explain our link analysis method based on diversity in detail. The proposed method is composed of two parts. The first is calculating the diversity of the set of pages. For this part, we propose a method to quantify how each page in the set is different from the others. This quantification method creates a feature vector of each page with LDA and then uses the sum of the distance between the centroid and each page in the set. The second part is expanding the HITS algorithm by using the diversity of the referrer and referral documents of each document. The method calculates the diversity of document links to the document as the "referrer diversity" and of the document linked by the document as the "referral diversity." We set these two diversity scores in the HITS algorithm as the weight of the edge.

The purpose of the proposed method is to find pages that are useful for everyone: not only specialists but novice users as well. We assume a simple HITS-based hypothesis: the Hub page that links to diverse Authorities and the Authority page that is linked by diverse Hubs must feature a wide readership and widespread interest and therefore be of general interest to everyone.

3.1 Determining Diversity

To calculate diversity, each document has to be expressed as a feature vector. In our method, we take topics from the main text of the document and define the diversity as how different the topics of a page set are. The most simple way to express the document as the vector is just counting all the terms in the main text, but when the number of documents increases, the number of vector dimensions explodes. We used LDA (Latent Dirichlet Allocation) to compress the dimension. Each term in the dataset is assumed to belong to one topic of i types of topics based on the topic model, where i is the given number of dimensions. LDA classifies terms into the topics to which they belong. The frequency of topic occurrence in each document is used as the feature vector of the document. Each document can be expressed by an i-dimensional vector. The length of documents in the dataset is not constant, so the feature vector has to be normalized.

Note that to create a feature vector, we can use other information in addition to the topic of the main text, such as the degree of confirmation or denial, the stance of the author, sentiments, and so on [5]. If another kind of diversity is needed, the algorithm can deal with it by switching function just as well based on a differently constructed feature vector.

We propose a diversity function $d(P)$ to calculate the diversity of the document set P. This function takes a high value when each document in P has a different topic and a low value when all documents have a similar topic.

Every document is defined as $n = (n_1, n_2, n_3, n_4, \ldots, n_k)$, that is, documents in the dataset are explained as a k dimension feature vector based on the frequency of the topic found in their main text. The diversity, which means how the documents in the set P are diverse, is defined as

$$d(P) = \frac{1}{|P|} \sum_{p \in P} dist(p, mean(P)), \tag{1}$$

where $mean(P)$ is the arithmetic mean of the set of documents P and $dist(a, b)$ is the Euclidean distance between vectors a and b. $mean(P)$ is

$$mean(P) = \frac{1}{|P|} \sum_{p \in P} p, \tag{2}$$

and $dist(a, b)$ is

$$dist(a, b) = \sqrt{\sum_{i=1}^{k} (a_i - b_i)^2}, \tag{3}$$

where a and b are vectors expressed as $a = (a_1, a_2, a_3, \ldots, a_k)$, $b = (b_1, b_2, b_3, \ldots, b_k)$. The definition of $d(P)$ in our method is the normalized sum of the difference between each page in P and its mean. When the dispersion of the documents in P is high, this value become high. The value drops into $[0, \frac{\sqrt{2}}{2}]$ when the norm of the feature vectors is normalized to 1. $d(P)$ gets its maximum value when all the articles in P have a different topic and gets its minimum value, 0, when all

documents in P have the same content or the number of documents in P is less than two. We call $d(In(n))$ the referrer diversity of n, which means how pages linking n are diverse, where $In(n)$ is the set of pages that link to n. Likewise, we call $d(Out(n))$ the referral diversity of n, which means how pages linked by page n are diverse, where $Out(n)$ is the set of pages linked by n. It is important to note that this calculus equation is not so novel. You can see the same equation in the K-means clustering as the value to minimize. It is used to express the cohesiveness of the cluster. When the dimension of the vector is one, this equation means the variance.

3.2 Diversity-Based HITS Algorithm

In this section, we explain the method to calculate the Hub score and the Authority score by using the diversity-based HITS algorithm considering referral diversity and referrer diversity.

First, we have to prepare the root set that is used for the result page of the given query. We also need a graph for the link analysis, so a base set was created as the sum set of the root set itself, linking document set and linked document set of the root set.

The original HITS algorithm defines Hubs and Authorities by mutual recursion, as

$$hub(p) = \sum_{q,p\to q} auth(q) \tag{4}$$

$$auth(p) = \sum_{q,q\to p} hub(q), \tag{5}$$

where both p and q is a page in the base set. It can be expressed as matrix calculation below:

$$h = Aa \tag{6}$$

$$a = A^T h, \tag{7}$$

where A is the adjacency matrix of the data set and h and a are the vectors of the Hubs and Authorities, respectively.

$$A_{ij} = \begin{cases} 1 \text{ if } i \text{ links } j \\ 0 \text{ otherwise.} \end{cases} \tag{8}$$

In this case, A_{ij} means the link between node j and node i.

Our method modifies this HITS calculation by using diversity-based factors $d(In(p))$ and $d(Out(p))$ as

$$dhub(p) = d(Out(p)) \sum_{q,p\to q} dauth(q) \tag{9}$$

$$dauth(p) = d(In(p)) \sum_{q,q\to p} dhub(q), \tag{10}$$

where $In(p)$ is the set of pages that links page p, and $Out(p)$ is the set of pages linked by page p. Two scores: $dhub(p)$ and $dauth(p)$ mean diversity-based Hubs and diversity-based Authorities. We call $d(In(p))$ as "referrer diversity" of page p, and $d(Out(p))$ as "referral diversity" of page p. This formula supports the two diversity-based hypotheses above, that is, "The Hub that links diverse Authorities is a good Hub" and "The Authority that is linked by diverse Hubs is a good Authority."

We can replace the adjacency matrix A of the HITS algorithm. We propose two diversity-based adjacency matrixes. One is based on referral diversity and the other on referrer diversity. The expanded adjacency matrix taking referral diversity is as below:

$$N_{ij} = \begin{cases} d(In(j)) & \text{if } i \text{ links } j \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where $In(j)$ is the set of pages that links page j. In this matrix, links to the document that are linked by many different documents are weighted highly and receive a higher score. In contrast, a higher score is not correlated with pages linked by many similar pages. The other diversity-based adjacency matrix, O, which takes the referral diversity, is

$$O_{ij} = \begin{cases} d(Out(i)) & \text{if } i \text{ links } j \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

where $Out(i)$ is the set of pages linked by page i. This matrix means that the weight of the link from the page linking diverse pages become high, and the weight of the link becomes low when the link is from a page linking similar pages. To consider referral diversity, A should be replaced with O. A can be replaced with N to consider referrer diversity.

$$dh = Oda \tag{13}$$

$$da = N^T dh, \tag{14}$$

where da and dh are the vectors of the diversity-based Hubs and diversity-based Authorities, respectively. In the original HITS algorithm, the coefficient of propagation is the same with the Hubs to Authorities propagation and the Authorities to Hubs propagation. The proposed method replaces A and A^T individually with diversity-based matrixes. It makes an asymmetric link weighted bipartite graph (see Fig.3). To calculate diversity-based HITS, it uses referral diversity score for back link propagation from Authority page to Hub page, and referrer diversity score for link propagation from Hub page to authority page.

The expanded HITS algorithm with a diversity-based adjacency matrix can be solved by the power method.

$$dh = ON^T dh \tag{15}$$

$$da = N^T Oda. \tag{16}$$

Vectors da and dh converge to Authorities and Hubs when they are normalized by every phase. Authority-based ranking can be used to find documents and Hub-based ranking can be used to find good linking sites or navigating sites.

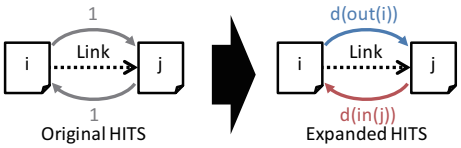

Fig. 3. Asymmetric propagation value.

4 Experiment

To compare the methods explained in section 3 with the original HITS algorithm and variant methods, we conducted an experimental Web ranking evaluation. The pages for the given query were sorted by the Authority score of each method as a search result ranking. Each of these rankings was evaluated by bucket-based evaluation. As stated previously, the aim of the proposed method is to enable both novice and expert users to find information that is useful. We used one participant who played a novice and classified documents sampled by each ranking into useful and useless documents after reading the main text.

4.1 Variant Methods

To clarify the effect of referral diversity and referrer diversity particularly, we have prepared two variant methods.

One is the referral diversity-based method. It replaces A with O, and A^T is unchanged. It considers only referral diversity and not referrer diversity. This supports the hypothesis that the "The Hub that links diverse Authorities is a good Hub."

Another one is the referrer diversity-based method. It replaces A^T with N^T, and A is unchanged. It considers only the referrer diversity. This supports the hypothesis that the "The Authority that is linked by diverse Hubs is a good Authority."

We compared them to proposed method as baseline methods.

4.2 Data Set

We used The ClueWeb09-JA Dataset, which contains over 67 million pages with 400 million links between them. We prepared eight queries, shown in Table 1. These included two types of query: those about difficult topics and those about easy topics. The top 1,000 pages on BM25 sorted ranking were extracted by each query as the root set. Pages linking to a page included in the root set and linked by pages in the root set were used as the base set. The root set was then sorted by each method. The page evaluated by the participant is a sample of the root set.

We compared four methods below.

- **Both** is the proposed method based on both diversity factors. Pages are ranked by the Authority score calculated by Eqn. 9. It uses the referrer diversity to calculate Authorities and the referral diversity to calculate Hubs.
- **Referrer** is the variant method based only on referrer diversity. It uses the referrer diversity factor on the propagation from Authorities to Hubs.
- **Referral** is the variant method based only on referral diversity. It uses the referral diversity factor on the propagation from Hubs to Authorities.
- **HITS** is the baseline method. It is the original HITS algorithm.

Each method is compared using the Authority-based score because in this experiment we want to find the document but not the linking page.

To compare rankings by these methods, we evaluated sample pages of each ranking. First, we split the ranking into 5 buckets: the top 10 pages and pages ranked from 11—50, 51—200, 201—500, and 501—1,000. We took 10 sample pages from each bucket. The sampling rate of each bucket was not constant: the upper part of the ranking was sampled in high density and the bottom part was sampled coarsely. All of the top 10 pages were evaluated by one participant, but only 2 % of the bottom pages were evaluated. The total number of evaluated pages was 1,366 by 8 queries and 5 methods after removing duplicates. The participant evaluated each document in terms of whether or not it was useful for a novice user. Sample pages were sorted randomly for every query. Each page in the dataset was shown as plain text.

We used GibbsLDA++[1] as an implementation of LDA. We classified the terms in the dataset into 100 topics. The LDA sampling was iterated 2,000 times.

Queries	Type
Postal service privatization	Easy
France trip	Easy
The Sagrada Familia	Easy
Fish called by different names in life stage	Easy
Parkinson's disease	Specialist
Game theory	Specialist
Compiler	Specialist
Machine learning	Specialist

Table 1. Queries.

4.3 Result

The results are shown in Table 2 and Fig. 4. The original HITS method ranked many correct pages in the middle of the ranking. All expanded methods were

[1] http://gibbslda.sourceforge.net/

influenced by the original method. In the ideal case, it is hoped that many correct pages appear in the top part of the ranking and that a small number of correct pages appear in the bottom. The **both** method, which uses both types of diversity, found more useful pages for novices in the top part of the ranking than the other methods. The **referrer** method seems a little bit better than the original HITS method in the top part of ranking, but it ranks many correct pages in the bottom part. The **referral** method had almost the same accuracy as the original HITS method. The ratio of correct pages in the data set was 0.22 through all queries.

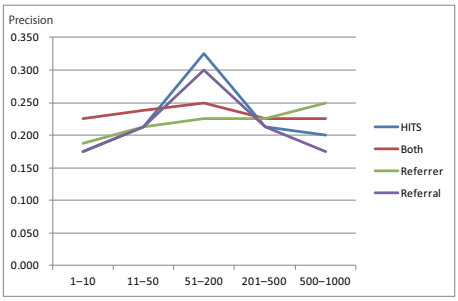

Fig. 4. Results for all queries.

Bucket	HITS	Both	Referrer	Referral
1–10	0.175	0.225	0.188	0.175
11–50	0.213	0.238	0.213	0.213
51–200	0.325	0.250	0.225	0.300
201–500	0.213	0.225	0.225	0.213
500–1000	0.200	0.225	0.250	0.175

Table 2. Precision of each bucket through all queries.

The eight queries used in the experiment can be separated into easy queries and specialist queries. Figure 5 shows the results for two types of query. In the easy query case, the ratio of correct pages is high: 0.35 through 4 queries. When the search task was easy, the search result contained useful pages for novice users. Each method had similar precision in the buckets of the upper part of the ranking. The **both** method found more correct pages than other methods in the middle part of the ranking. The **referral** method had findings throughout the ranking. The **referrer** method performed worse than the original HITS. In the specialist query case, the total number of correct pages in the dataset was small, with a ratio of just 0.09. It is assumed that pages on specialist topics are commonly not written in a language easy enough for novice users to understand.

On the whole, the **both** method, which uses both referral and referrer diversity, works well, especially when the query is specialist. The HITS algorithm ranked correct pages in the middle part of the ranking, and the two **referral** or **referrer** methods used the original adjacency matrix A. They were influenced strongly by original HITS result. Although the total number of correct pages was small, the **both** method ranked many correct pages in the top part of the ranking and not many of them in the bottom.

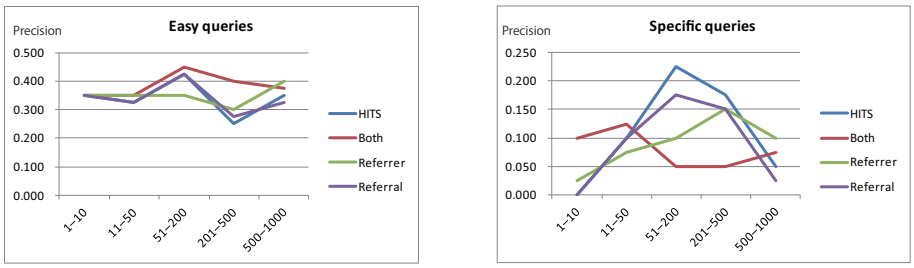

Fig. 5. Results by two types of queries.

5 Discussion

The proposed method works better when the query is specific. For example, when the query is "machine learning," the proposed method finds three suitable documents while basic HITS can not find any suitable documents. The detail of pages judged as correct document are online dictionary sites, or introductory works written by academic society. Dictionary sites and encyclopedia sites are frequently linked by many individual personal blogs. The author of each blog has different interests. Authors that are incidentally interested in "machine learning" will create links to those pages. Another day, they write about their interests, and link other interested pages. Then these pages have links not only to "machine learning" pages.

On the other hand, pages judged as not suitable are documents deemed too difficult, low-quality pages, spam and pages not relevant to the query. In this case, some social bookmarking sites were found in the top part of the basic HITS ranking. They do not contain useful information. These sites strongly connected with themselves by internal links. They are characterized by a high score for the both Hubs and Authority. The basic HITS algorithm is weak to such kind of link structures. Pages from social bookmarking services feature a similar design template. Our method estimated their topics as similar to each other.

The proposed method did not work well when the query was easy. The number of correct pages found is same to HITS in the top part of ranking. Our method found more correct page in the middle and bottom part of ranking. On this task, the number of pages suitable for novice users is inherently large. When the query is general, relevant pages are general too. For instance, when the query was "France trip," each method yielded results mostly containing content from major travel agency sites, all of which are about hotels, touristic hot spots or itineraries. When a judgment is determined only by the relevancy between the query and page, HITS-based algorithms may not be suitable even if it was expanded. There is a possibility that the diversity factor couses reverse effect, that is, the document linked only by documents about a trip may be relevant to the trip. Then, non-diverse tight links provide higher relevancy.

The graph of basic HITS algorithm has its peak in the middle of the ranking, where it scores good pages. In the bottom of ranking, there are many pages

linked by few pages. Most of these pages are spam pages, low-quality pages, or minor pages. In the top part of ranking, a lot of individual pages of major online service sites appeared. They are linking to each other. Some of them have no content inside, i.e., private pages in social bookmarking websites, message pages in online fora with the aim to drive communication, product introduction pages of big company websites and so on. These pages are perceived as spam, or are not relevant to the query.

Even though our method works better than original HITS algorithm in this exeriment, it's accuracy is not high enough yet. The method has to be optimized by fixing and tuning parameters. For instance, the method used in this expetiment are not tuned, so it did not take into account the weight of propagation score and diversity score, the distribution of diversity score, the tuning of LDA and so on.

6 Conclusion

We proposed a diversity-based Web ranking method that expands on the HITS algorithm to include two diversity-based hypothesis: 1) that a page linking diverse Authorities is a valuable Hub, and 2) that a page linked by diverse Hubs is a valuable Authority. The objective of the diversity-based HITS algorithm is to enable not limited specialist users but general users to find suitable documents, that is, documents that are useful for novice users. We defined diversity as how the topic of each document in a set of documents is different from the topics of the other documents. We call the diversity of pages linking to the page referrer diversity and the diversity of pages linked by the page referral diversity. We expanded the HITS algorithm by replacing the adjacency matrix with two diversity-based matrixes. The proposed methods were compared with the original HITS algorithm by their authority scores in terms of finding useful pages for novice users. The method that uses both referral and referrer diversity could rank more good pages high, especially when the search query was specific.

As future work, we intend to expand the diversity-based methods further. Our method abandoned many factors to simplify the model. Of course, the method itself is built around the idea of diversity, not popularity, so it is necessary to focus on the number of linking documents, the amount of information, and the power of influence pages. Moreover, diversity can be defined not only from the topic of pages: for example, we can define it as authors' property, sentiment of documents, temporal-spatial metadata, and so on. We will tackle these issues with additional diversity-based methods.

Acknowledgments

This work was supported in part by the following projects: Grants-in-Aid for Scientific Research (Nos. 24240013) from MEXT of Japan and by JSPS Fellows ($24 \cdot 5417$).

References

1. Surowiecki, J.: The wisdom of crowds. Anchor (2005)
2. Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '98, New York, NY, USA, ACM (1998) 335–336
3. Wang, J., Zhu, J.: Portfolio theory of information retrieval. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. SIGIR '09, New York, NY, USA, ACM (2009) 115–122
4. Capannini, G., Nardini, F.M., Perego, R., Silvestri, F.: Efficient diversification of web search results. Proc. VLDB Endow. **4**(7) (April 2011) 451–459
5. Minack, E., Demartini, G., Nejdl, W.: Current approaches to search result diversification. In: Proceedings of The First International Workshop on Living Web at the 8th International Semantic Web Conference (ISWC). (Oct. 2009)
6. Stirling, A.: A general framework for analysing diversity in science, technology and society. Journal of the Royal Society Interface **4**(15) (2007) 707–719
7. Haveliwala, T.: Topic-sensitive pagerank: a context-sensitive ranking algorithm for web search. Knowledge and Data Engineering, IEEE Transactions on **15**(4) (july-aug. 2003) 784 – 796
8. Gyöngyi, Z., Garcia-Molina, H., Pedersen, J.: Combating web spam with trustrank. In: Proceedings of the Thirtieth international conference on Very large data bases - Volume 30. VLDB '04, VLDB Endowment (2004) 576–587
9. Takahashi, Y., Ohshima, H., Yamamoto, M., Iwasaki, H., Oyama, S., Tanaka, K.: Evaluating significance of historical entities based on tempo-spatial impacts analysis using wikipedia link structure. In: Proceedings of the 22nd ACM conference on Hypertext and hypermedia. HT '11, New York, NY, USA, ACM (2011) 83–92
10. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM **46**(5) (September 1999) 604–632
11. Deng, H., Lyu, M.R., King, I.: A generalized co-hits algorithm and its application to bipartite graphs. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD '09, New York, NY, USA, ACM (2009) 239–248
12. Lempel, R., Moran, S.: The stochastic approach for link-structure analysis (salsa) and the tkc effect. Computer Networks **33**(1 – 6) (2000) 387–401
13. Tong, H.: Fast random walk with restart and its applications. In: In ICDM ' 06: Proceedings of the 6th IEEE International Conference on Data Mining, IEEE Computer Society (2006) 613–622
14. Nakatani, M., Jatowt, A., Ohshima, H., Tanaka, K.: Quality evaluation of search results by typicality and speciality of terms extracted from wikipedia. In: Proceedings of the 14th International Conference on Database Systems for Advanced Applications. DASFAA '09, Berlin, Heidelberg, Springer-Verlag (2009) 570–584
15. Akamatsu, K., Pattanasri, N., Jatowt, A., Tanaka, K.: Measuring comprehensibility of web pages based on link analysis. In: Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01. WI-IAT '11, Washington, DC, USA, IEEE Computer Society (2011) 40–46