Analysis and Improvement of HITS Algorithm for Detecting Web Communities

Saeko Nomura Satoshi Oyama Tetsuo Hayamizu
Toru Ishida
Department of Social Informatics, Kyoto University
Honmachi Yoshida Sakyo-ku, Kyoto, 606-8501 Japan
{saeko, oyama, hayamizu}@kuis.kyoto-u.ac.jp, ishida@i.kyoto-u.ac.jp

Abstract

In this paper, we discuss problems with HITS (Hyperlink-Induced Topic Search) algorithm, which capitalizes on hyperlinks to extract topic-bound communities of web pages. Despite its theoretically sound foundations, we observed HITS algorithm failed in real applications. In order to understand this problem, we developed a visualization tool LinkViewer, which graphically presents the extraction process. This tool helped reveal that a large and densely linked set of unrelated Web pages in the base set impeded the extraction. These pages were obtained when the root set was expanded into the base set. As remedies for this topic drift problem, prior studies applied textual analysis method. On the other hand, we propose two methods which utilize only the structural information of the Web: 1) The projection method, which projects eigenvectors on the root subspace, so that most elements in the root set will be relevant to the original topic, and 2) The base-set downsizing method, which filters out the pages without links to multiple pages in the root set. These methods are shown to be robust for broader types of topics and low in computation cost.

1. Introduction

The World Wide Web continues to expand in size and complexity: The number of Web pages is estimated to grow up to 100 billion within two years [4] and the number of domain names is reported to be 93,047,785 worldwide as of July 2000 [1]. As the size increases, its complexity grows so enormously that we can no longer grasp the whole picture. However, within small, local areas, the Web is still structured orderly because the link structure is built upon a considerable effort of human annotation. Authors of web pages tend to make links to other pages on related topics. By making use of these links, we can extract and group pages relevant to the topics. In this paper, we call these groups of pages as "Web communities" ([10], [8]).

Among the algorithms proposed for this purpose, HITS (Hyperlink-Induced Topic Search) algorithm is studied most widely. This algorithm models communities as interconnection between 'authorities' and 'hubs.' Despite the theoretical foundations, HITS algorithm is reported to fail in some real situations. In this paper, we discuss those problems and propose some remedies.

In responding to the problems, prior studies suggest taking into account the content of pages (i.e., word frequencies). However powerful, this approach undermines HITS algorithm's theoretical motivation to extract relevant pages based solely on link structure. In contrast, this study improved the algorithm using structural information more effectively. Our approach consists of three steps.

- We developed a tool LinkViewer to figure out how HITS algorithm failed. LinkViewer visualizes the page extraction process as well as the results.
- Observing HITS algorithm with LinkViewer, we found out what kind of problems occurred at which stage.
- Without using the content analysis, we modified HITS algorithm only by the hyperlink information of the Web.

The paper is structured as follows. Section two reviews the HITS algorithm and describes our implementation. Section three introduces LinkViewer, a visualization tool for HITS algorithm, and discusses problems of the algorithm. Section four proposes algorithms that resolve problems detected in the previous section, and Section five further discusses our algorithms compared to those in prior studies.

2. HITS (Hyperlink-Induced Topic Search) Algorithm

2.1. An overview of the HITS algorithm

HITS algorithm mines the link structure of the Web and discovers the thematically related Web communities that

consist of 'authorities' and 'hubs.' Authorities are the central Web pages in the context of particular query topics. For a wide range of topics, the strongest authorities consciously do not link to one another. Thus, they can only be connected by an intermediate layer of relatively anonymous hub pages, which link in a correlated way to a thematically related set of authorities [9]. These two types of Web pages are extracted by iteration that consists of following two operations.

$$x_p = \sum_{q,q \to p} y_q$$

$$y_p = \sum_{q,p \to q} x_q$$

For a page p, the weight of x_p is updated to be the sum of y_q over all pages q that link to p: where the notation $q \to p$ indicates that q links to p. In a strictly dual fashion, the weight of y_p is updated to be to the sum of x_q. Therefore, authorities and hubs exhibit what could be called mutually reinforcing relationships: a good hub points to many good authorities, and a good authority is pointed to by many good hubs. The whole picture of HITS algorithm is shown in Figure 1.

step 1: Collect the r highest-ranked pages for the query σ from a text-based search engine such as AltaVista[2]. These r pages are referred as the root set $R\sigma$.

step 2: Obtain the base set $S\sigma$ whose size is n, by expanding $R\sigma$ to include any page pointed to by pages in $R\sigma$ and at most d pages pointing to pages in $R\sigma$.

step 3: Let G[$S\sigma$] denote the subgraph induced on the pages in $S\sigma$. Two types of links in G [$S\sigma$] are distinguished as *transverse links* and *intrinsic links*. The former are the links between pages with different domain names, and the latter are the ones between pages with the same domain name. All *intrinsic links* from the graph $S\sigma$ are deleted, keeping only the edges corresponding to *transverse links.*

step 4: Make the n by n adjacency matrix A and its transposed matrix A^T. Normalized principal eigenvector $\boldsymbol{e}_1$ of $A^T A$ that corresponds to the largest eigenvalue λ_1 is obtained by eigenvalue calculation.

step 5: Find elements with large absolute values in the normalized principal eigenvector $\boldsymbol{e}_1$. Return them as 'authorities.'

(note: Kleinberg[9] set each parameter as follows: r=200 and d=50. As a result, n becomes about 1,000 to 5,000.)

Figure 1. HITS algorithm.

2.2. Problems with the HITS Algorithm

To clarify problems with HITS algorithm, we traced Kleinberg's experiments. We picked 9 query topics for our study: 'abortion,' 'Artificial Intelligence,' 'censorship,' 'Harvard,' 'jaguar,' 'Kyoto University,' 'Olympic,' 'search engine,' and 'Toyota.' In these query topics, all but 'Kyoto University' and 'Toyota' were used in [9] and [8]. Though we fixed the parameters r, d, and a text-based search engine for collecting the root set to examine Kleinberg's experiments rigorously, we observed HITS algorithm performed poorly in several of our test cases. In this paper, we discuss focusing on topic 'Artificial Intelligence' as a successful example, and topic 'Harvard'as an unsuccessful example.

Topic: 'Artificial Intelligence'

The extracted top 5 authorities and hubs of 'Artificial Intelligence' in our experiment are indicated in Table 1. The decimal fractions shown on the left of URLs represent authority weights (x_p) and hub weights (y_p) respectively.

Table 1. Authorities and hubs of 'Artificial Intelligence.'

x_p	Authorities
.372	http://www.cs.washington.edu/research/jair/home.html
.298	http://www.aaai.org/
.294	http://www.ai.mit.edu/
.272	http://ai.iit.nrc.ca/ai_point.html
.234	http://sigart.acm.org/
y_p	Hubs
.228	http://yonezaki-www.cs.titech.ac.jp/member/hidekazu/Work/AI.html
.228	http://www.cs.berkeley.edu/~russell/ai.html
.204	http://uscia1.usc.clu.edu/pantonio/cco360/AIWeb.htm
.181	http://www.scms.rgu.ac.uk/staff/asga/ai.html
.171	http://www.ex.ac.uk/ESE/IT/ai.html

(note: x_p and y_p represent authority weight and hub weight respectively.)

The top authority was the home page of JAIR (Journal of Artificial Intelligence Research), the second authority was AAAI (American Association for Artificial Intelligence), then MIT AI laboratory followed. Namely, famous organizations related to Artificial Intelligence based in the United States were successfully extracted. This AI community was supplemented by hubs, which consisted of the researcher's personal Web pages (e.g. S. Russell at UCB).

Topic: 'Harvard'

In Kleinberg's experiment, authorities of 'Harvard' were related to Harvard University; e.g. the homepage of Harvard University, Harvard Law School, Harvard Business School,

and so on. However, in our experiment, the Web pages authored by a financial consulting company were extracted (see Table 2). These pages did not relate to query 'Harvard.'

Table 2. Authorities and hubs of 'Harvard.'

x_p	Authorities
.130	http://www.wetradefutures.com/investment.asp
.130	http://www.wetradefutures.com/trend.htm
.130	http://www.wetradefutures.com/market technology.htm
.130	http://www.wetradefutures.com/florida investment.htm
.130	http://www.wetradefutures.com/investing investment.htm
y_p	Hubs
.247	http://www.profittaker.net/data.htm
.247	http://www.profittaker.org/new twentyseven.htm
.247	http://profittaker.com/sunday trader more.htm
.247	http://www.profittaker.cc/system software.htm
.247	http://www.futureforecasts.com/contact phone.htm

(note: x_p and y_p represent authority weight and hub weight respectively.)

In this case, higher ranked 56 authorities had the same authority weights, and higher ranked 5 hubs had the same hub weights. By checking the contents of these pages, we detected that these authorities and hubs were authored by a single organization.

3. Visualizing the Problems

To observe HITS algorithm's behavior, we developed a visualization tool named *LinkViewer*. In this section, we explain the function of LinkViewer, and find out what kind of problems occur at which stage of HITS algorithm.

3.1. HITS Algorithm Visualizing Tool: *LinkViewer*

LinkViewer is software to visualize the convergence process of HITS algorism. When the iterative procedure converges, the topology of extracted Web community is figured out. According to the Web browsing function of this tool, we can confirm the page contents of extracted authorities and hubs.

Figure 2 shows the user interface of LinkViewer. Users input and output the base set files from file menu above the image, then command iteration using a computing button. By operating a playback box located below the image, users can see the convergence process of Web communities.

Nodes (Web pages) displayed in the window of LinkViewer are colored in red and blue. Red nodes represent authorities and blue nodes represent hubs. Higher ranked authorities/hubs are colored deeper. In the box expressing the gradations, which is located in the left bottom of the image, the number of nodes for each rank of authority/hub is indicated.

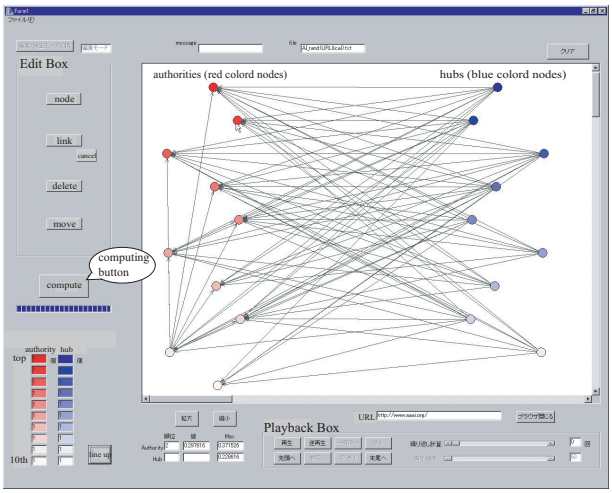

Figure 2. Userinterface of LinkViewer: a successful example 'Artificial Intelligence.'

3.2. Identifying the Problems

By comparing the results of 'Artificial Intelligence' and 'Harvard,' we found a large difference in the topology of Web communities between them.

With respect to the topic 'Artificial Intelligence' (see Figure 2), good hubs pointed to many good authorities and good authorities were pointed to by many good hubs.

On the other hand, in the case of 'Harvard,' numerous pages had the same authority weights and hub weights, and 87 pages densely linked each other (See the upper cluster of Figure 3). Although we could find the home page of Harvard University as the 67th authority on the LinkViewer, pages linking with this page disappeared in the process of iteration. The cluster below the home page of Harvard University was a CBS sports line archive. These Web pages were not related to the query topic 'Harvard' either.

The cluster of 87 pages that consists of higher ranked authorities and hubs was obtained through a single Web page (the 66th authority), which was included in the root set. When the root set was expanded into the base set, 86 other pages densely linking with this single page were obtained. This problem was observed in the rest of unsuccessful topics either.

In brief, this problem is expressed as Figure 4. Since pages in the root set contain query terms, they tend to have relative information to the original topics. According to the characteristics of Web communities ([10], [8]), there is every possibility that these pages link to each other. Therefore, a group of pages represented by dots in Figure 4 ought to be extracted as a topically related Web community.

On the other hand, pages which do not link to multi-

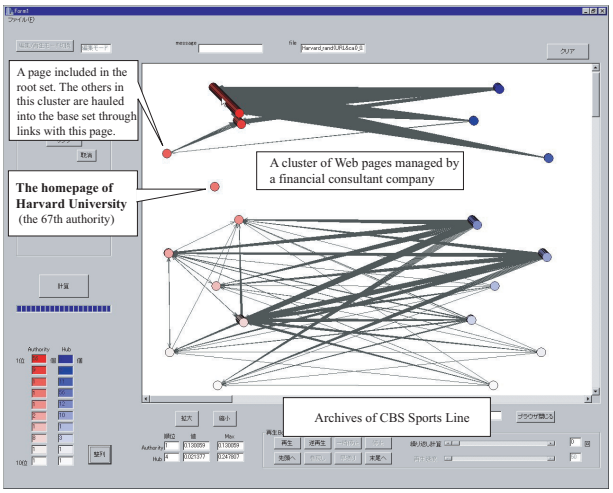

Figure 3. Userinterface of LinkViewer: an unsuccessful example 'Harvard.'

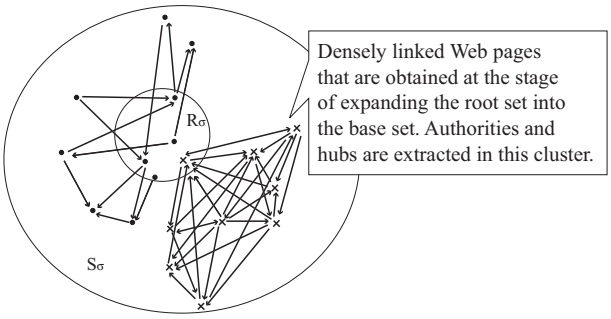

Figure 4. An unsuccessful phenomenon of HITS algorithm.

ple pages in the root set tend to be irrelevant to the original query topics (pages represented by crosses in Figure 4). However, if they link very densely each other, they are extracted as authorities and hubs against users' will. These are correspond to the pages authored by a financial consultant company and CBS news on the query topic 'Harvard.'

According to our analysis, the problem with HITS algorithm can be summarized as follows.

> **The problem with HITS Algorithm**
> In the base set, a considerable number of irrelevant Web pages to the original topics are included. If these pages densely link each other, HITS algorithm converge into them and cannot extract appropriate Web communities in accordance with the original query topics.

4. Improvement of HITS Algorithm

The phenomenon that authorities converge into densely linked irrelevant pages is called *topic drift problem*. This problem is notorious in the area of Information Retrieval. To address this problem, we propose two types of link-analysis-based modification: *the projection method* and *the base-set downsizing method*. These methods take into account the number of links to/from pages included in the root set to extract appropriate Web communities.

4.1. The Projection Method

The projection method modifies the HITS algorithm at the stage of eigenvalues computation.

In the first place, we try to interpret the matrix $A^T A$. The (i, j) element of matrix $A^T A$ is expressed as the following equation.

$$\begin{aligned}(A^T A)_{ij} &= \sum_{k=1}^{n} (A^T)_{ik}(A)_{kj} \\ &= \sum_{k=1}^{n} (A)_{ki}(A)_{kj}\end{aligned}$$

Since $(A)_{ij}$ indicates the citation from page i to page j, $(A^T A)_{ij}$ represents the number of common links to page i and page j. In short, $(A^T A)_{ij}$ is a similarity matrix among the Web pages, with the similarity measured by the number of common links to page i and page j.

Now, let V be the $n-$dimensional Euclidean space corresponding to the base set, and W the $r-$dimensional subspace corresponding to the root set. We refer to the latter as the root subspace.

All eigenvalues of $A^T A$ are arranged in decreasing order $\lambda_1, \lambda_2, \ldots, \lambda_n$. Corresponding to these eigenvalues, normalized eigenvectors are referred as $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$. Then, the matrix $A^T A$ can be decomposed as below, on the assumption that each eigenvalue differs.

$$A^T A = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^T + \lambda_2 \boldsymbol{e}_2 \boldsymbol{e}_2^T + \ldots \lambda_n \boldsymbol{e}_n \boldsymbol{e}_n^T$$

A case where authorities do not converge into original query topics is illustrated in Figure 5. (In Figure 5, although W and the orthogonal complement W are drawn in a single dimension, they contain multiple dimensions practically.)

In Figure 5, while λ_1 is larger than λ_2 ($\lambda_1 > \lambda_2$), most elements of $\boldsymbol{e}_1$ are in the orthogonal complement of W ($W^{\perp}$), which is added at the stage of expanding the root set into the base set. This means many common links from pages not included in the root set (the group of pages which is expressed as $\times$ in Figure 4). On the other hand, $\boldsymbol{e}_2$ contains many elements in the root subspace. This means that

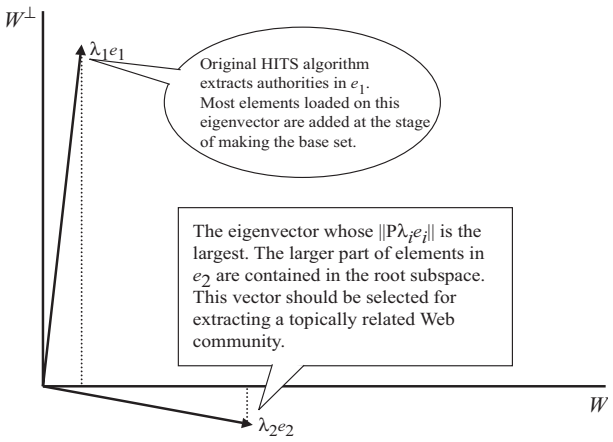

Figure 5. The vector e_2 which ought to be selected to extract appropriate Web community.

Web pages indicated by e_2 have a large number of common links from the root set (the group of pages that are expressed as dots in Figure 4). To obtain e_2 as a vector to extract authorities, we apply projection P from V to W. The inequality $\|P\lambda_1 e_1\| < \|P\lambda_2 e_2\|$ holds, e_2 can be utilized for extracting authorities.

According to the observations above, we improve HITS algorithm as Figure 6.

step 4 (a): Make the n by n adjacency matrix A and its transposed matrix A^T. Compute all normalized eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$, that correspond to each eigenvalue $\lambda_1, \lambda_2, \ldots, \lambda_n$, by iteration.

(b): Apply the projection P on the root subset W for each eigenvector and find the eigenvector $\boldsymbol{e}^*$ which maximize the value of $\|P\lambda_i \boldsymbol{e}_i\|$.

step 5: Find elements with large absolute values in the normalized eigenvector $\boldsymbol{e}^*$. Return them as 'authorities.'

Figure 6. The projection method.

The point in this argument is that we utilize the common links from the pages in the root set to estimate the topically relatedness. This means the target pages for extracting authorities are throughout the base set. Therefore, if numerous pages which belong to W densely link to/from a small number of pages which belong to $W^{\perp}$, pages in $W^{\perp}$ (which do not contain original query terms) have possibility to be extracted as authorities. On the other hand, if the root set is not expanded into the base set, there is no possibility for these pages to be authorities.

In this method, however, since we have to compute all eigenvalues and eigenvectors to extract appropriate authorities, computational cost is large. Therefore, the iterative method computes eigenvalues in decreasing order, so that we can set criteria for terminating iteration. Let λ^* and $\boldsymbol{e}^*$ be a pair of eigenvalue and eigenvector which maximizes $\|P\lambda_i \boldsymbol{e}_i\|$ of all computed so far. Since $\|P\boldsymbol{e}_i\| \leq 1$ holds, subsequent pairs do not exceed the current $\|P\lambda^* \boldsymbol{e}^*\|$ when once λ_i becomes smaller than $\|P\lambda^* \boldsymbol{e}^*\|$. We can stop iterations at this point and the number of eigenvalues and eigenvectors necessary to be computed are curtailed.

4.2. The Base-set Downsizing Method

In our experiments, the size of base set turned out to be from 4,500 to 8,200. These base sets contain a large number of unrelated Web pages. Since the Web communities that should be extracted have dense links to/from the pages in the root set, we perform filtering at when obtaining the base set as Figure 7. This method can curtail the eigenvalue computational cost efficiently.

step 2 (a): Obtain the base set $S\sigma$ whose size is n, by expanding $R\sigma$ to include any page pointed to by pages in $R\sigma$ and at most d pages pointing to pages in $R\sigma$.

(b): *Among the pages obtained at the stage of expanding* $S\sigma$, extract any pages pointed to by *multiple pages* and any page that points to *multiple pages* in $R\sigma$, and make *the downsized base set*.

Figure 7. The base-set downsizing method.

The computational cost at the stage of downsizing the base set is estimated as follows. Let us consider the n by n adjacency matrix of $S\sigma$. Removing pages which link to multiple pages in $R\sigma$ means that removing rows whose sum of r elements is two and over from the $n - r$ rows. Therefore, needed computational cost here is $r(n - r)$. With respect to the pages being linked to by multiple pages in $R\sigma$ need the same amount of computational cost.

In fact, the size of the base set could be reduced to less than 1/10 by this base-set downsizing method in all query topics. Since the cost of eigenvector computation is n^3[11], we could cut the computational cost drastically, even with $2r(n - r)$ extra filtering cost.

4.3. The Integration of The Two Algorithms

The base-set downsizing method eliminates pages that have less than k links to/from the pages in the root set. In this paper, we set k=2. On the one hand, if k is set smaller, we cannot resolve the topic drift problem in more topics.

On the other hand, if k is set larger, more pages that should be included in the Web communities might be eliminated. In this way, the base-set downsizing method has a difficulty in setting the appropriate filtering strength.

Therefore, we integrate the projection method to the result of the base-set downsizing method (k=2). This means that we eliminate obviously unrelated pages in the first place, then apply projection P from downsized Euclidian space corresponding to the downsized base-set V' to the root subspace W. Then, even if the appropriate Web communities for the query topic is loaded on the non-principal vectors, we can find them.

4.4. Result of HITS Algorithm Improvements

In this subsection, we report the result of our improvements applied to topic 'Harvard.'

The left side of Table 3 shows the result of 'Harvard,' to which the projection method was applied. The top authority became the home page of Harvard University, followed by pages of Harvard University Library Resources, Harvard Medical Web, Harvard Online Gateway, and so on. In short, we succeeded to obtain the appropriate Web community related to 'Harvard' by the projection method. These pages are loaded on the negative end of third non-principal vector.

Table 3. Extracted top five authorities of 'Harvard' by the projection method and the base-set downsizing method.

The projection method		The base-set downsizing method	
.907	http://www.harvard.edu	.852	http://www.harvard.edu
.138	http://hplus.harvard.edu	.174	http://hplus.harvard.edu
.116	http://www.med.harvard.edu	.151	http://www.med.harvard.edu
.102	http://web.mit.edu	.132	http://www.fas.harvard.edu
.088	http://www.haa.harvard.edu	.127	http://www.haa.harvard.edu

(note: The decimal fractions represent authority weights (x_p).)

As shown in Table 4, the computation of $||P\lambda_i \boldsymbol{e}_i||$ was curtailed at i=4 and e^* became $\boldsymbol{e}_3$. The principal eigenvector produced authorities relating to a financial consultant corporation, which original HITS algorithm extracted. In regard to in non-principal second vector, a CBS sports line archive became authorities.

With respect to the result of the base-set downsizing method is shown on the right side in Table 3. Similarly to the projection method, we succeeded to obtain a group of pages relating to Harvard University. Here, since the size of base set scaled down from 8,209 pages to 815 pages, we could cut big computational cost.

Table 4. The computation of $||P\lambda_i \boldsymbol{e}_i||$ (curtailed at i=4).

i	λ_i	$\|\|P\lambda_i \boldsymbol{e}_i\|\|$
1	1034.654419	14.594835
2	689.295044	4.583474
3	624.957703	597.531805
4	531.576416	–

Although the base-set downsizing method could solve the topic drift problem on this topic, we applied the projection method to this result. As a result, the same authorities were extracted on the principal eigenvector.

5. Analysis and Discussion

In this section, we evaluate the proposed algorithms applying topics that the original algorithm failed to handle. For successful topics ('Artificial Intelligence', 'search engine' and 'Toyota'), the proposed improvements could produce the same results as the original one.

This section discusses the novel point of our methods compared to prior related studies either.

5.1. Evaluation of Proposed Methods

Table 5 at the last page of this article shows that the projection method succeeds to extract appropriate communities in all topics. A topic 'Kyoto University' is a striking example. While the original algorithm falsely picked up mathematics departments in universities throughout the world, improved algorithm by the projection method could distil departments of 'Kyoto University.' Yet, these communities of pages are loaded not on the principal eigenvectors but on lower ones.

In the topic 'abortion,' the pro-life community, such as NRLC (National Right to Life), ALL (American Life League), HLI (Human Life International) and so on come to the front. With regard to the topic 'censorship,' 'jaguar,' and 'Olympic,' the advocacy of freedom of speech community, automobile community, and Olympic games community are extracted respectively.

The base-set downsizing method improves but still fails in some topics. As indicated by '$\times$' in Table 5, irrelevant pages are extracted for topics 'abortion' and 'Kyoto University.' These results show the difficulty in appropriate filtering strength settings of the base-set downsizing method. This means that we could not eliminate unrelated pages at the threshold k=2. With respect to topic 'Kyoto University' for example, some pages abstracted by AltaVista were some

kinds of university index in Japan. Since the Web pages that linked to these pages had multiple links with other pages in the root set, the topic drift problem could not be resolved.

Integration of these two improvement methods gives better results than either of them. Not only can relevant pages be extracted but also those pages are loaded on principal eigenvectors or the non-principal vectors in low computation cost. In particular, for the topics in which base-set downsizing method could alone succeed, pages are loaded on principal eigenvectors (See the lowest line of each topic in Table 5). To make sure that our results differed from the keyword search results on AltaVista, we indicate the example of topic 'abortion' and 'censorship' in Table 6.

Table 6. Top five results of query 'abortion' and 'censorship' on AltaVista.

'abortion'
http://www.abortion.com/
http://www.abortion-help.com/
http://www.prochoice.org/
http://www.abortiontv.com/
http://www.abortioninfo.net/
'censorship'
http://www.indexoncensorship.org/
http://www.journalismnet.com/media/censorship.htm
http://newslinx.internet.com/newstopics/censorship.html
http://www.censorshipkills.com/
http://www.liberty.org.uk/

As Kleinberg argues [9], HITS algorithm extracts multiple densely linked collections of hubs and authorities on multiple eigenvectors. The projection method proposed in this paper rearranges multiple eigenvectors produced by HITS algorithm based on the hyperlink information. Therefore, this improvement does not harm the characteristics of HITS algorithm. For example, with respect to the topic 'jaguar,' not only the community of automobile, but also the community of video games (on the 13th non-principal vector) were extracted in our experiment.

5.2. Related Work

Prior studies (such as [5], [6], [3], [7]) that attempted to improve HITS algorithm can be classified into either "hyperlink analysis method" or "integration method of hyperlink and textual analyses." In this context, our approach is placed in the first category.

To address *the topic drift problem*, some studies proposed the integration method. Bharat and Henzinger prune all pages whose *relevance weights* are below a specified threshold from the base set [3]. The *relevance weight* is computed based on the match between query terms and phrases in a Web page. In [5], authority/hub weights of each page are computed based on the match between query terms and anchor texts in source pages.

On the other hand, the hyperlink analysis method is applied only to particular problems such as mutually reinforcing relationships between hosts caused by duplicated pages (i.e., mirror sites) [3], [7]. This approach reduces authority weights and hub weights as follows. If there are m links from pages on a first host to a single page on a second host, they give each link an authority weight of $1/m$. The same procedure is conducted for computing hub weights as well.

We proposed the method that ranks eigenvectors using the similarity to the original query topics by counting the links to/from pages in the root set. This approach solves the topic drift problem by coping with multiple problems, not only individual problem such as defusing the mutually reinforcing relationships between hosts.

6. Conclusion

In this paper, we addressed the problem with HITS algorithm. We developed a visualizing tool *LinkViewer* to figure out how HITS algorithm failed. By observing the behavior of the algorithm with LinkViewer, we found out following problems: In the base set, when a large and densely linked set of unrelated Web pages exists, HITS algorithm converge into them and cannot extract appropriate Web communities.

While other studies integrated textual analysis into hyperlink analysis as remedies for this *topic drift problem*, we modified it only by the "link-only" approach as follows.

1) **The projection method**, which projects eigenvectors on the root subspace so that most elements in the root set will be relevant to the original topic.

   When computing eigenvalues and normalized eigenvectors of the matrix $A^T A$, elements with large absolute value are extracted as authorities in eigenvector $\boldsymbol{e}^*$. Here, $\boldsymbol{e}^*$ has the largest $\|P\lambda_i \boldsymbol{e}_i\|$ (P means a projection to the root subspace W). To curtail computational costs, computations are closed λ_i becomes smaller than $\|P\lambda^* \boldsymbol{e}^*\|$.

2) **The base-set downsizing method**, which filters out the pages without links to multiple pages in the root set.

   In $S\sigma$, pages that have links to/from k pages in $R\sigma$ are extracted. For this downsized base set, iteration is conducted to extract authorities and hubs.

Though the base-set downsizing method can cut the computational cost effectively, it has a difficulty in setting the filtering strength k. Therefore, we proposed the integration of the two improvement methods. This method gives

good results. Not only can relevant pages be extracted but also those pages are loaded on principal eigenvectors or the non-principal vectors in low computation cost for broader types of topics.

In future we plan to explore better estimates of multiple communities for a single query topic. For example, on query topic 'jaguar,' though we could find automobile community on the non-principal 4th vector very easily, there exist some more different communities on different eigenvectors (e.g. video game community on the 13th vector) that we could not extract automatically. Developing a tool, which mines multiple communities and visualize the topology of them will be our next task.

Acknowledgement

The work is supported by the Japan Science Technology Corporation for Core Research for Evolutional Science and Technology Project "Universal Design of Digital City."

References

[1] http://www.isc.org/ds/.

[2] http://www.altavista.com.

[3] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Research and Development in Information Retrieval*, pages 104–111, 1998.

[4] D. Butler. Souped-up search engines. *Nature*, 405:112–115, May, 2000.

[5] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proc. 7th International World Wide Web Conference*, Brisbane, Australia, 1998.

[6] S. Chakrabarti, B. Dom, R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the Web's link structure. *IEEE Computer*, 32(8):60–67, 1999.

[7] J. Dean and M. Henzinger. Finding related pages in the World Wide Web. In *Proc. 8th International World Wide Web Conference*, Toronto, Canada, 1999.

[8] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring Web communities from link topology. In *Proc. 9th ACM Conference on Hypertext and Hypermedia (HyperText 98)*, pages 225–234, Pittsburgh PA, June 1998.

[9] J. Kleinberg. Authoritative sources in a hyperlinked environment, 1997. Research Report RJ 10076 (91892), IBM.

[10] J. Kleinberg. Hubs, Authorities, and Communities. *ACM Computing Surveys*, 31(4es, Article No.5), 1999.

[11] W. A. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in C: the art of scientific computing(2nd ed.)*. Cambridge University Press, New York, 1997.

Table 5. Comparison among extracted authorities: The original HITS algorithm and proposed improvements.

	'abortion'	'censorship'	'Kyoto University'
The original HITS algorithm	http://www.amazon.com/exec/obidos/redirect-home/youdebatecom/	http://www.lycos.com/network/	http://www.kusm.kyoto-u.ac.jp/
	http://www.amazon.com/exec/obidos/	http://www.lycos.com/network/find/	http://www.math.tohoku.ac.jp/
	http://rd1.hitbox.com/	http://www.lycos.com/network/shop/	http://www.maths.monash.edu.au/
	http://www.amazon.com/exec/	http://www.lycos.com/lycosinc/	http://phoenix.levels.unisa.edu.au/
	http://www.amazon.com/	http://hotwired.lycos.com/home/	http://www.ma.adfa.oz.au/
The projection method	*4th non-principal vector*	*6th non-principal vector*	*7th non-principal vector*
	http://www.nrlc.org/	http://www.eff.org	http://www.kyoto-u.ac.jp/
	http://www.all.org/	http://www.cdt.org/	http://www.kogaku.kyoto-u.ac.jp/
	http://www.hli.org/	http://www.aclu.org/	http://www.kais.kyoto-u.ac.jp/
	http://www.prolife.org/ultimate/	http://www.eff.org/blueribbon.html	http://www.media.kyoto-u.ac.jp/
	http://www.priestsforlife.org/	http://www.epic.org/	http://www.educ.kyoto-u.ac.jp/
The base-set downsizing method	× http://stocks.uscity.net/	http://www.eff.org/	http://www.kyoto-u.ac.jp/
	× http://weather.uscity.net/	http://www.ncac.org/	http://www.kyokyo-u.ac.jp/
	× http://auction.uscity.net/	http://www.aclu.org/	http://www.kufs.ac.jp/
	× http://chat.uscity.net/	http://www.cdt.org/	http://www.kyoto-phu.ac.jp/
	× http://forums.uscity.net/	http://www.eff.org/blueribbon.html	× http://www.u-tokyo.ac.jp/
The integration of the two improvements	*2nd non-principal vector*	*principal eigenvector*	*2nd non-principal vector*
	http://www.nrlc.org/	http://www.eff.org/	http://www.kurims.kyoto-u.ac.jp/
	http://www.plannedparenthood.org/	http://www.ncac.org/	http://www.media.kyoto-u.ac.jp/
	http://www.naral.org/	http://www.aclu.org/	http://www.rri.kyoto-u.ac.jp/
	http://www.prochoice.org/	http://www.cdt.org/	http://www.kudpc.kyoto-u.ac.jp/
	http://www.gynpages.com/	http://www.eff.org/blueribbon.html	http://www.kogaku.kyoto-u.ac.jp/

	'jaguar'	'Olympic'
The original HITS algorithm	http://www.chelmerfineart.com/	http://www.gannett.com/
	http://www.chelmerfineart.com/adam_barsby1.htm	http://autofinder.cincinnati.com/
	http://www.chelmerfineart.com/adam_barsby10.htm	http://careerfinder.cincinnati.com/
	http://www.chelmerfineart.com/adam_barsby2.htm	http://homefinder.cincinnati.com/
	http://www.chelmerfineart.com/adam_barsby3.htm	http://cincinnati.com/search/
The projection method	*non-principal 4th vector*	*4th non-principal vector*
	http://www.jaguarcars.com/	http://www.olympic.org/
	http://www.jag-lovers.org/	http://www.sydney.olympic.org/
	http://www.jagweb.com/	http://www.olympic-usa.org/
	http://www.jec.org.uk/	http://www.slc2002.org/
	http://www.jags.org/	http://www.australian.olympic.org.au/
The base-set downsizing method	http://www.jaguarcars.com/	http://www.olympic.org/
	http://www.jag-lovers.org/	http://www.sydney.olympic.org/
	http://www.jagweb.com/	http://www.olympic-usa.org/
	http://www.jec.org.uk/	http://www.australian.olympic.org.au/
	http://www.collection.co.uk/	http://www.slc2002.org/
The integration of the two improvements	*principal eigenvector*	*principal eigenvector*
	http://www.jaguarcars.com/	http://www.olympic.org/
	http://www.jag-lovers.org/	http://www.sydney.olympic.org/
	http://www.jagweb.com/	http://www.olympic-usa.org/
	http://www.jec.org.uk/	http://www.australian.olympic.org.au/
	http://www.collection.co.uk/	http://www.slc2002.org/

(note: '×' means irrelevant Web pages to the query topics.)