Constructing Good Quality Web Page Communities

Jingyu Hou
Department of Mathematics and Computing
University of Southern Queensland
Toowoomba, Qld 4350, Australia
jingyu@usq.edu.au

Yanchun Zhang
School of Computing
University of Tasmania
Hobart, Tasmania 7001, Australia
yan@utas.edu.au

Abstract

The World Wide Web is a rich source of information and continues to expand in size and complexity. To capture the features of the Web at a higher level to realise the information classification and efficient retrieval on the Web is becoming a challenge task. One natural way is to exploit the linkage information among the Web pages. Previous work such as HITS in this area is based on a set of retrieved pages to get a Web community that is a bunch of pages related to the query topics. Since the set of retrieved pages may contain many unrelated pages (noise pages) to the query topics, the obtained Web community sometimes is unsatisfactory. In this paper, we propose an innovative algorithm to eliminate noise pages from the set of retrieved pages and improve its quality. This improvement will enable existing community construction algorithms to construct good quality Web page communities. The proposed algorithm reveals and takes advantage of the relationships among concerned Web pages at a deeper level. The numerical experiment results show the effectiveness and feasibility of the algorithm. This algorithm could also be used solely to filter unnecessary Web pages and reduce the management cost and burden of Web-based data management systems. The ideas in the algorithm can also be applied to other hyperlink analysis.

Keywords: World Wide Web, Web community, hyperlink analysis, singular value decomposition (SVD).

1 Introduction

Nowadays, with rapid expanding in size and complexity, the World Wide Web is a rich source of information to be explored. To capture the features of the Web at a higher level to realise the information classification and efficient retrieval on the Web is becoming a challenge. Traditional Web information retrieval is based on the keywords and implemented by Web search engines, such as *Yahoo!*, *AltaVista*. The search result is a set of Web pages that may or may not contain the required information. On the other hand, users cannot identify, from this retrieved result, which pages are relevant or at a high relevant rank to their query topics. It is always difficult for users to go further to search relevant pages and get the required information except browsing listed pages one by one. This is not the desirable case.

To tackle this problem, one approach is to re-organize or classify the obtained pages into different groups that are relevant to the given query topics to a certain extent. These groups form a Web page community. For this purpose, relationships among the concerned pages should be revealed. One natural way to achieve this goal is to exploit the linkage information among them. This is mainly because the Web is in hypertext linkage. Another reason is that, in most cases, authors of the Web pages create links to other pages with an idea in mind that the linked pages are relevant to the linking pages. Therefore, the linkage relationships among the Web pages more or less reveal their mutual semantic connections. The representative work in this area is the HITS (Hyperlink-Induced Topic Search) algorithm proposed by Kleinberg (1999). This algorithm tries to construct Web communities from a set of retrieved Web pages by analysing their hyperlinks and by iterative operations. The Web communities consist of *authorities* and *hubs*. An authority page is a page that contains the most definitive, central and useful information in the context of particular query topics. A hub page is a page that points to many of the authorities (Kleinberg 1999). The authorities and hubs exhibit mutually reinforce relationship: a good hub points to many good authorities; a good authority is pointed to by many good hubs. The goal of the HITS algorithm is to obtain the Web communities with good authorities and hubs.

Other related works (e.g. Chakrabarti, Dom, Gibson, Kleinberg, Raghavan, and Rajagopalan 1998, Bharat and Henzinger 1998) improve this algorithm by combining page content analysis techniques and graph edge weighting. As these works observed, the set of retrieved pages contains many pages that are unrelated to the query topics (i.e. contain no query terms). If these pages are in high linkage density, they will dominate the iteration operations and the obtained authority or hub pages may be not relevant to the query topics, which is called *topic drift* problem. We call these unrelated pages as *noise pages*. Actually, given a query topic, a Web browser could retrieve and return a set of Web pages that are considered to be the most related to the query topic by the browser. This initial set of retrieved pages is called *root set*. The root set of pages could then be extended to form a new set of pages by adding more pages to the root set. These added pages have linkage relationships with the pages in the root set. We call this extended set of root set as *base set* (details of root and base set construction are described in section 2). The Web community construction algorithms are based on this base set of pages. Therefore, the quality of the base set of pages, i.e. the percentage of

topic-related pages in the base set of pages, mainly determines the quality of the produced Web community. However, in the procedure of extending a root set to a base set of pages, many topic-unrelated (noise) pages would be added to the base set. Previous works (e.g. Chakrabarti, Dom, Gibson, Kleinberg, Raghavan, and Rajagopalan 1998, Bharat and Henzinger 1998, Dean and Henzinger 1999) mainly consider reducing the influence of these noise pages to the community construction algorithms *after* the base set has been established.

In this paper, we will propose an innovative algorithm to eliminate noise pages from the base set of pages *B* and get another good quality base set of pages *B'*, from which a good quality Web community can be constructed (see figure 1). This algorithm purely takes advantage of the link relationships among the pages in *B*. To be precise, the algorithm considers the link relationships between the pages in root set *R* and pages in *B*-*R*. Here, *B*-*R* is a page set and a page in it belongs to *B* but does not belong to *R*. These link relationships are expressed in a linkage matrix *A*. By the singular value decomposition of the matrix *A* (Hou, Zhang, Cao, Lai and Ross 2001, Hou, Zhang, Cao and Lai 2000), the relationship between pages at a deeper level will be revealed, and a numerical threshold could be defined to eliminate noise pages (see figure 1). This approach is based on a reasonable assumption that the pages in the root set are topic-relevant and noise pages are mainly brought in by the procedure of expanding root set *R* to base set *B*. Indeed, the root set *R* may also contain noise pages, though the possibility is less. However, by eliminating noise pages from the base set, the influence of the remained noise pages in root set will be greatly reduced and good quality communities could be obtained.

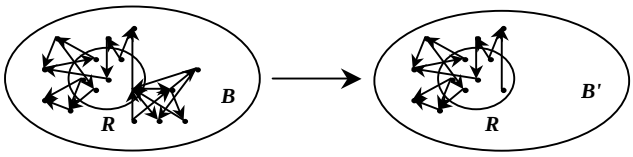

Figure 1: Getting new base set with less noise pages by applying the proposed algorithms

This paper is organized as follows. In section 2, some background about the HITS algorithm and its improvements are given for better understanding our current work. In section 3, the algorithm for eliminating noise pages from the base set of pages is presented. The algorithm is based on the singular value decomposition (SVD) of a matrix. Therefore, some background about SVD is also provided in this section. In section 4, some numerical experiment results and their analysis are provided to show the effectiveness and feasibility of the proposed algorithm. Some related work is discussed in section 5. Finally, conclusions and further research directions are presented in section 6. The algorithm depiction is listed in the appendix of this paper.

2 HITS Algorithm Background

HITS algorithm was proposed originally in 1997 and formally in 1999 by Kleinberg (1999). It is based on the assumption that if document (page) *A* has a hyperlink to document *B*, then the author of document *A* thinks that document *B* contains valuable information. Therefore, it is possible to use the in-degree[1] of a document to measure its quality. On the other hand, if a document *A* points to a lot of good documents, then *A*'s opinion becomes more valuable, and the fact that *A* points to *B* would suggest that *B* is a good document. The goal of the HITS algorithm is to exploit the linkage information between Web pages and get the Web community. This algorithm actually considers the mutual influence between the pages, rather than simply counting the number of links for each page.

The HITS algorithm consists of three main procedures:

1. Collecting *r* highest-ranked pages for the user-supplied query σ from a text-based search engine (e.g. *AltaVista*) to form the root set of pages *R*. Growing *R* to form the base set of pages, *B*, by adding to *R* more pages which are pointed by or pointing to the pages in *R*. *B* is considered to be a query specific graph whose nodes are pages.
2. Associating with each page *p* in *B* a hub weight *h(p)* and an authority weight *a(p)* with initial values of 1. Then iteratively updating the *h(p)* and *a(p)* ($p \in B$) according to the following iterative operations:

$$a(p) = \sum_{\substack{q \to p \\ q \in B, q \neq p}} h(q), \qquad h(p) = \sum_{\substack{p \to q \\ q \in B, q \neq p}} a(q)$$

   in which "$p \to q$" denotes "page *p* has a hyperlink to page *q*". Normalize the vectors *a* and *h* after each iteration.
3. After iteration reaches steady point (i.e. values of vectors *a* and *h* will not change any more), abstracting *s* pages (as authorities) with *s* highest *a( )* values together with the *s* pages (as hubs) with the *s* highest *h( )* values to be the core of a community.

Kleinberg (1999) proved that vectors *a* and *h* converge. Thus the termination of the iteration is guaranteed. From our numerical experiment experience, with the absolute error precision 10^{-4}, the number of iteration is 20.

When the above HITS algorithm is applied to the base set of pages, it does not work well in many cases. Usually, it meets the following problems (Bharat and Henzinger 1998): mutually reinforcing relationships between *hosts*, automatically generated links, and non-relevant nodes. Mutually reinforcing relationships between hosts occur when a set of pages on the first host point to a single page on the second host, or one page on the first host points to multiply pages on the second host. This will greatly and unreasonably increases the impact of one host to the Web community construction. Automatically generated links are produced by Web page generating tools. These links usually do not represent a human's opinion. The non-relevant nodes may cause the *topic drift* problem.

[1] In-degree of a document is the number of documents that link to this document.

To tackle the first problem, (Bharat and Henzinger 1998) improved the algorithm by assigning authority or hub weights to the edges of graph *B*. If there are *k* edges from documents on the first host to a single document on the second host, each edge will be given an authority weight (*auth_wt*) of *1/k*. If there are *l* edges from a single document on the first host to a set of documents on the second host, each edge will be given a hub weight (*hub_wt*) of *1/l*. Then the iterative operation in the HITS algorithm is improved as

$$a(p) = \sum_{\substack{q \to p \\ q \in B, q \neq p}} h(q) \times auth\_wt(q, p),$$

$$h(p) = \sum_{\substack{p \to q \\ q \in B, q \neq p}} a(q) \times hub\_wt(p, q).$$

Bharat and Henzinger (1998) also proved that the vectors *a* and *h* converge, and the termination of the iteration can be guaranteed.

There is also other work in improving the HITS algorithm to tackle other problems by combining the structural analysis (linkage analysis) with the page content analysis. For example, in (Bharat and Henzinger 1998), in order to eliminate the automatically generated links and non-relevant nodes, a similarity measurement is introduced exploiting the content of the pages. In (Chakrabarti, Dom, Gibson, Kleinberg, Raghavan, and Rajagopalan 1998), page content analysis is used to increase the linkage weight between two pages. Since we only consider using the pure structural analysis to improve the quality of Web communities, we will not give more details about content analysis here.

3 Algorithm for Eliminating Noise Pages

As indicated in the above algorithms, the base set of pages is the base for constructing the Web community. Therefore, its quality has a great influence to that of communities. However, the previous work mainly concerns about how to reduce the influence of the noise pages. From another view angle, if most noise pages in the base set can be filtered or eliminated before the community construction algorithm is applied, the quality of communities would be greatly increased. This is the point from which our algorithm is developed.

It can be seen from the HITS algorithm that the base set of pages is derived from the root set of pages by adding more pages in it. This procedure will bring more query topic related pages into the base set, but bring many topic unrelated pages into the base set as well. For example, for the query topic (term) "*Harvard*", apart from many pages about Harvard University in the base set of pages, there are also many other pages in it that do not contain this query term, such as the page for a beer company (*http://www.johnsonbeer.com/*) and the page for a comedy club (*http://www.punchline.com/*), due to their links to some pages in the root set. In order to eliminate these noise pages, we can reasonably assume that the root set of pages is topic-related. According to our numerical experiment and analysis (see section 4 of this paper), if an authority page or a hub page is a topic-drift one, it is usually located in those pages that connect to a small part of root set (fewer connections) but link densely with each other. They dominated the algorithm operation and caused the topic drift problem. Such pages should be recognised as noise pages and eliminated. On the other hand, if a page has fewer connections with the root set, it is most likely to be a topic unrelated page (noise page) and could not be included in the base set in most cases. However, another question has arisen. What is the threshold for "*fewer connections*"? This problem cannot be solved only by directly counting the number of links for each page. It should be solved by considering the mutual influence between the pages in the base set and defining an exact threshold. Our algorithm is trying to reveal the relationships among the concerned pages at a deeper level, and precisely define the threshold for eliminating noise pages by exploiting this revealed relationship. In this algorithm, the linkage information between the pages is directly expressed as a matrix, from which deeper relationships between these pages are revealed by some matrix operations. Thanks to the singular value decomposition (SVD) of a matrix in linear algebra that can reveal the internal relationship between matrix elements (Deerwester, Dumais, Furnas, Landauer and Harshman 1990) (Hou, Zhang, Cao, Lai and Ross 2001) (Hou, Zhang, Cao and Lai 2000), we apply it to this situation and propose our algorithm based on it. For better understanding our algorithm, some background knowledge of SVD is provided as follow.

3.1 Singular Value Decomposition (SVD) Background

The SVD definition of a matrix is as follow:

Let $A = [a_{ij}]_{m \times n}$ be a real $m \times n$ matrix. Without loss of generality, we suppose $m \geq n$ and the rank of A is $rank(A) = r$. Then there exist orthogonal matrices $U_{m \times m}$ and $V_{n \times n}$ such that

$$A = U \begin{pmatrix} \Sigma_1 \\ 0 \end{pmatrix} V^T = U \Sigma V^T \qquad (1)$$

where $U^T U = I_m, V^T V = I_n, \Sigma_1 = diag(\sigma_1, ..., \sigma_n)$, $\sigma_i \geq \sigma_{i+1} > 0$ for $1 \leq i \leq r-1$, $\sigma_j = 0$ for $j \geq r+1$, Σ is a $m \times n$ matrix, U^T and V^T are the transpositions of matrices U and V respectively, I_m and I_n represent $m \times m$ and $n \times n$ identity matrices separately. The *rank* of A indicates the maximal number of independent rows or columns of A. Equation (1) is called the singular value decomposition of A. The singular values of A are defined as the diagonal elements of Σ (i.e. $\sigma_1, \sigma_2, ..., \sigma_n$). The columns of U are called left singular vectors and those of V are called right singular vectors (Datta 1995, Golub and Van Loan 1993).

The SVD could be used effectively to extract certain important properties relating to the structure of a matrix, such as the number of independent columns or rows,

eigenvalues, approximation matrix and so on. Since the singular values of A are in non-increasing order, it is possible to choose a proper parameter k such that the last r-k singular values are much smaller than the first k singular values, and these k singular values dominate the decomposition. The next theorem reveals this fact.

Theorem [Eckart and Young]. Let the SVD of A be given by equation (1) and $U = [u_1, u_2, \ldots, u_m]$, $V = [v_1, v_2, \ldots, v_n]$ with $0 < r = rank(A) \le \min(m, n)$, where u_i, $1 \le i \le m$ is an m-vector, v_j, $1 \le j \le n$ is an n-vector and

$$\sigma_1 \ge \sigma_2 \ge ... \ge \sigma_r > \sigma_{r+1} = ... = \sigma_n = 0 .$$

Let $k \le r$ and define

$$A_k = \sum_{i=1}^{k} u_i \cdot \sigma_i \cdot v_i^T . \qquad (2)$$

Then

1. $rank(A_k) = k$;
2. $\min_{rank(B)=k} \| A - B \|_F^2 = \| A - A_k \|_F^2 = \sigma_{k+1}^2 + ... + \sigma_r^2$
3. $\min_{rank(B)=k} \| A - B \|_2 = \| A - A_k \|_2 = \sigma_{k+1}$.

Where $\| A \|_F^2 = \sum_{j=1}^{n} \sum_{i=1}^{m} | a_{ij} |^2$ and $\| A \|_2^2 = max$ (*eigenvalues of* $A^T A$) are measurements of matrix A.

The proof can be found in (Datta 1995). This theorem indicates that matrix A_k, which is constructed from partial singular values and vectors (see Figure 2), is the best approximation to A (i.e. conclusions 2 and 3 of the theorem) with rank k (conclusion 1 of the theorem).

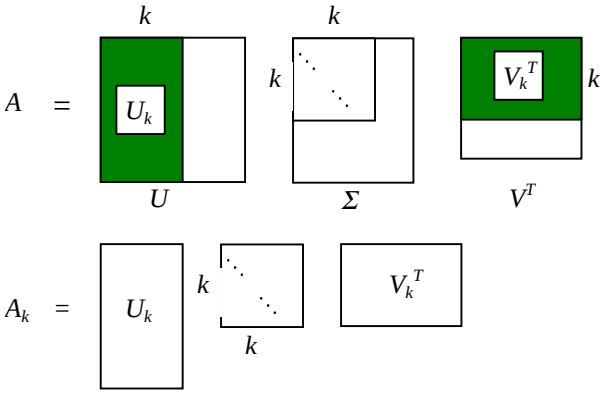

Figure 2: Construction of approximation matrix A_k

The above theorem suggests that for a proper parameter k, A_k is the best approximation of A, and the difference between them is very small. In other words, A_k captures the main structure information of A and minor factors in A are filtered. Our algorithm is just taking advantage of this important property to reveal the deeper relationship between pages from their linkage matrix, and define the threshold for eliminating noise pages. In practical computation, since $k \le r$ and only partial matrix elements are involved, matrix computation cost could be reduced if k is chosen properly. The next sub-section gives details of our algorithm.

3.2 Noise Page Elimination Algorithm (NPEA)

When the base set of pages is constructed for the user's query, the linkage information among the pages is also obtained. There are two types of links to be distinguished, *transverse links* and *intrinsic links*. The transverse links are the links between pages with different domain names[2], and the intrinsic links are the links between pages with the same domain names. Since intrinsic links very often exist purely to allow for infrastructure navigation of a site, they convey much less information than transverse links about the authority of the pages they point to (Kleinberg 1999). As in (Kleinberg 1999), intrinsic links in our algorithm are deleted from the obtained links and only the transverse links are kept. We denote the root set of pages R as a directed graph $G(R)=(R,E_R)$: the nodes correspond to the pages, and a directed edge $(p,q) \in E_R$ indicates a link from p to q. Similarly, the base set of pages B is denoted as a directed graph $G(B)=(B,E_B)$. From the construction procedure of B, it can be easily inferred that $R \subset B$ and $E_R \subset E_B$.

Suppose the size of R (the number of pages in R) is n and the size of B is m. For the pages in R, the linkage matrix $S = (s_{ij})_{n \times n}$ could be constructed as

$$s_{ij} = \begin{cases} 1 & when\ (i, j) \in E_R \ \ or \ \ (j, i) \in E_R \ \ or \ \ i = j \\ 0 & otherwise. \end{cases}$$

It represents the link relationships between the pages in R. For the pages in B-R, another linkage matrix $A = (a_{ij})_{(m-n) \times n}$ for page $i \in (B\text{-}R)$ and page $j \in R$ could also be constructed as

$$a_{ij} = \begin{cases} 1 & when\ (i, j) \in E_B - E_R \ \ or \ \ (j, i) \in E_B - E_R \\ 0 & otherwise. \end{cases}$$

This matrix directly represents the linkage information between the pages in the root set and those not in the root set. The ith row of the matrix A, which is an n-dimensional vector, could be viewed as the coordinate vector of the page i in an n-dimensional space S_R spaned by the n pages in R.

For any two vectors $v1$ and $v2$ in an n-dimensional space S_n, as known in linear algebra, their similarity (or closeness) can be measured by their inner product (dot product) in S_n. The elements in $v1$ and $v2$ are the coordinates of $v1$ and $v2$ in the S_n respectively. In the page set B-R, since each page is represented as an n-dimensional vector (a row of matrix A) in the space S_R, all the similarities between any two pages in B-R can be expressed as AA^T. On the other hand, as indicated in sub-section 3.1, there exists a SVD for the matrix A:

$$A_{(m-n) \times n} = U_{(m-n) \times (m-n)} \Sigma_{(m-n) \times n} V^T_{n \times n} .$$

Therefore, the matrix AA^T can also be expressed as

$$AA^T = (U\Sigma)(U\Sigma)^T .$$

[2] Domain name here means the first level of the URL string associated with a Web page.

From this equation, It is obvious that matrix $U\Sigma$ is equivalent to the matrix A, and the ith (i = 1,…, m-n) row of matrix $U\Sigma$ could be naturally and reasonably viewed as the coordinate vector of the page i (*page* $i \in$ *B-R*) in another n-dimensional space S'_R. Similarly, for the matrix S, there exists a SVD of S:

$$S_{n\times n} = W_{n\times n}\Omega_{n\times n}X^T_{n\times n}.$$

The ith (i = 1,…, n) row of matrix $W\Omega$ is viewed as the coordinate vector of the page i (*page* $i \in R$) in another n-dimensional space S''_R.

For the SVD of matrix A, the matrix U could be expressed as $U_{(m-n)\times(m-n)} = [u_1, u_2, ..., u_{m-n}]_{(m-n)\times(m-n)}$ where u_i (i = 1,…,m-n) is a m-n dimensional vector $u_i = (u_{1,i}, u_{2,i}, ..., u_{m-n,i})^T$, and matrix V as $V_{n\times n} = [v_1, v_2, ..., v_n]_{n\times n}$ where v_i (i = 1,…, n) is an n dimensional vector $v_i = (v_{1,i}, v_{2,i}, ..., v_{n,i})^T$. Suppose *rank(A)* = r and the singular values of matrix A are

$$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > \sigma_{r+1} = ... = \sigma_n = 0.$$

For a given threshold δ ($0 < \delta \leq 1$), we choose a parameter k such that

$$(\sigma_k - \sigma_{k+1}) / \sigma_k \geq \delta,$$

and denote

$$U_k = [u_1, u_2, ..., u_k]_{(m-n)\times k},\ V_k = [v_1, v_2, ..., v_k]_{n\times k},$$
$$\Sigma_k = diag(\sigma_1, \sigma_2, ..., \sigma_k).$$

Let

$$A_k = U_k\Sigma_k V_k^T.$$

As the theorem in sub-section 3.1 indicates, A_k is the best approximation to A with rank k. Accordingly, the ith row R_i of the matrix $U_k\Sigma_k$ is chosen as the coordinate vector of page i (*page* $i \in$ *B-R*) in a k-dimensional subspace of S'_R:

$$R_i = (u_{i1}\sigma_1, u_{i2}\sigma_2, ..., u_{ik}\sigma_k), \quad i = 1, 2, \ldots, m\text{-}n. \qquad (3)$$

Since matrix A contains linkage information between the pages in *B-R* and R, from the properties of SVD and choice of parameter k, it can be inferred that coordinate vector (3) captures the main linkage information between the page i in *B-R* and the pages in R. The extent to which main linkage information is captured depends on the value of parameter δ. The greater the value of δ is, the more minor linkage information is captured. From the procedure of SVD (Datta 1995, Golub and Van Loan 1993), coordinate vector transformation (3) refers to linkage information of every page in *B-R*, and whether a linkage in matrix A is dense or sparse is determined by all pages in *B-R*, not just by a certain page. Therefore, equation (3) reflects mutual influence of all the pages in *B-R* and reveals their relationships at a deeper level. This situation is similar to those in (Deerwester, Dumais, Furnas, Landauer and Harshman 1990) (Hou, Zhang, Cao, Lai and Ross 2001) and (Hou, Zhang, Cao and Lai 2000).

In a similar way, suppose *rank(S)*=t and the singular values of matrix S are

$$\omega_1 \geq \omega_2 \geq ... \geq \omega_t > \omega_{t+1} = ... = \omega_n = 0.$$

The ith row R'_i of the matrix $W_t\Omega_t$ is chosen as the coordinate vector of the page i (*page* $i \in R$) in a t-dimensional subspace of S''_R:

$$R'_i = (w_{i1}\omega_1, w_{i2}\omega_2, ..., w_{it}\omega_t), \quad i = 1, 2, \ldots, n. \qquad (4)$$

Without loss of generality, let $k = min(k,t)$. The vector R_i can be expanded from a k-dimensional subspace to a t-dimensional subspace as

$$R_i = (u_{i1}\sigma_1, u_{i2}\sigma_2, ..., u_{ik}\sigma_k, \underbrace{0,0,...,0}_{t-k}), \qquad (5)$$
$$i = 1, 2, \ldots, m\text{-}n.$$

In order to compare the closeness between a page in *B-R* and the root set R, we project each page i in *B-R* (i.e. vector R_i of (5)) into the n-dimensional space spaned by the pages in R (i.e. vectors R'_i of (4), i = 1,…, n). The projection of page i (*page* $i \in$ *B-R*) PR_i is defined as

$$PR_i = (PR_{i,1}, PR_{i,2}, ..., PR_{i,n}),\ i = 1, 2, \ldots, m\text{-}n, \qquad (6)$$

where

$$PR_{i,j} = (R_i, R'_j) / \| R'_j \|$$
$$= \Big(\sum_{k=1}^{t} R_{ik} \times R'_{jk}\Big) \Big/ \Big(\sum_{k=1}^{t} {R'_{jk}}^2\Big)^{1/2},\ j = 1, 2, \ldots, n.$$

Within the same space, which is spaned by the pages in R, it is possible to compare the closeness between a page in *B-R* and the root set R. In other words, a threshold for eliminating noise pages can be defined. In fact, for each PR_i, if

$$\| PR_i \| = \Big(\sum_{j=1}^{n} PR_{i,j}^2\Big)^{1/2} \geq c_{avg}, \qquad (7)$$

where

$$c_{avg} = \sum_{j=1}^{n} \| R'_j \| \Big/ n,$$

then the page i in *B-R* could be remained in the base set of pages B. Otherwise, it should be eliminated from B. The parameter c_{avg} in the above equation represents the average link density of the root set R, and is the representative measurement for root set R. It is used as a threshold for eliminating the noise pages. Intuitively, if a page in *B-R* is a most likely noise page, it usually has fewer links with the pages in R. Thus its measurement $\|PR_i\|$ in (7) would be small and it is most likely to be eliminated. It is obvious that another representative measurement for root set R can also be defined as an elimination threshold. For example, the parameter c_{avg} could be replaced by $c_{\max} = \max_{j\in[1,n]}(\| R'_j \|)$ or $c_{\min} = \min_{j\in[1,n]}(\| R'_j \|)$. We call the algorithm with

parameters $c_{avg}, c_{\max}, c_{\min}$ as *avgAlgo*, *maxAlgo* and *minAlgo* respectively. Theoretically, the *avgAlgo* is ideal for elimination in most cases. The *maxAlgo* sometimes is too strict and many topic-related pages may be eliminated from the *B*-*R*. The *minAlgo* in some cases is too loose to eliminate many noise pages. In the next section, we will examine their numerical experiment results in eliminating noise pages and see if the experiment results are coincident with this theoretical analysis.

The above noise page elimination algorithm is depicted as the algorithm *NPEA* and listed in Appendix of this paper.

4 Numerical Experiment Results

In this section, we apply our algorithms to a situation where the original HITS algorithm fails to get satisfactory results in our experiment. This situation is for a query term "*Harvard*". The root set of pages, which are considered to be relevant to this term, is returned by a text-based Web search engine *AltaVista*. The construction of base set *B* is the same as that in (Kleinberg 1999) or in section 2. The size of *B* (the number of pages in *B*) is 8064, and the size of root set *R* is 200. We will firstly examine the numerical results of three algorithms (*avgAlgo*, *maxAlgo*, *minAlgo*) in noise page elimination with different values of parameter δ. From the analysis of these numerical results and our experimental experience, we will suggest which algorithm and parameter value are suitable in most cases. Meanwhile, we will show, via numerical results, that our algorithm enables the topic-related pages to capture the main linkage information among the concerned pages. Secondly, we will apply HITS algorithm to two situations and get two sets of authorities and hubs in order to see our algorithm really improves the quality of the Web community. One situation is that the noise pages in *B* are not eliminated; another situation is that the noise pages in *B* are eliminated by our algorithm.

In order to understand the experiment results better, we give the following definitions.

- *Suspected pages* are those pages that are topic-related but have at most one link to the pages in root set *R*.
- *Noise Page Filtering Rate* (NPFR) = number of filtered noise pages / total number of noise pages.
- *Noise Page Filtering Percentage* (NPFP) = number of filtered noise pages / total number of filtered pages.
- *Suspected Page Filtering Percentage* (SPFP) = number of filtered suspected pages / total number of filtered pages.
- *Efficient Filtering Percentage* (EFP) = NPFP + SPFP.

One important concept to be clarified for these definitions is what page is noise page. Here the noise page is in the meaning of common sense, i.e. it contains no query terms. In our experiment, the number of noise pages is 2968. Suspected pages are defined to distinguish those pages that are most likely to be *noise pages for the HITS algorithm*, but are not noise pages in common sense, as we stated before that noise pages usually have fewer links to the root set *R*. For example, in our experiment, the page "*http://www.hugo-sachs.de*" contains query term "*Harvard*", but it only have one link to the pages in the root set and have many links with a set of pages that contain no query term "*Harvard*" and produce an topic-drift authority page (see table 6) "*http://www.biochrom.co.uk/biochrom.htm*". In this case, the page "*http://www.hugo-sachs.de*" is a suspected page. Therefore, the efficient filtering percentage (EFP) reflects the highest percentage of filtered noise pages (for HITS) in all filtered pages (i.e. if all the suspected pages are noise pages for HITS).

Table 1 shows the numerical results of three algorithms (*avgAlgo*, *maxAlgo*, *minAlgo*) in noise page elimination with different value ranges of parameter δ. In our experiment, within each value range ($\delta \geq 0.4$ or $\delta \leq 0.3$), the number of filtered pages changes slightly with the changes of the δ value in that range. For simplicity, these minor changes are ignored in this table. From this table, it can be seen that the greater the value of parameter δ is, the less pages are eliminated (filtered). This is because with greater δ value, more minor linkage information is included in the coordinate vector of each page (equation (3)), thus the measurement of each page (equation (7)) is increased and number of filtered pages is decreased. These numerical results are coincident with the theoretical analysis in section 3.

Within the first value range of parameter δ ($\delta \geq 0.4$), although the noise page filtering rate (NPFR) of *maxAlgo* is 100%, its efficient filtering percentage (EFP) is only 71%. That means this algorithm eliminated too many topic related pages at the same time of eliminating all noise pages. This is not an ideal situation. For another two algorithms *avgAlgo* and *minAlgo*, although their efficient filtering percentages (EFPs) are the same (94%), the noise page filtering rate (NPFR) of *avgAlgo* (98%) is much better than that of *minAlgo* (81%). These numerical results show that within the range of $\delta \geq 0.4$, *avgAlgo* is an ideal noise page elimination algorithm.

For the second value range of δ ($\delta \leq 0.3$), similar to the above analysis, *maxAlgo* is not an ideal algorithm either. Although the noise page filtering rates (NPFRs) of *avgAlgo* and *minAlgo* are the same (98%), the efficient filtering percentage (EFP) of *minAlgo* (92%) is better than that of *avgAlgo* (87%). So in this case, the *minAlgo* is an ideal algorithm for this experiment.

The above numerical results and analysis indicate that *maxAlgo* is not suitable for noise page elimination because it eliminates too many topic related pages at the same time. It seems that with small δ value ($\delta \leq 0.3$), *minAlgo* should be adopted for noise pages elimination; with large δ value ($\delta \geq 0.4$), *avgAlgo* should be adopted. But in this experiment, we found the page linkage distribution within the root set *R* is relatively even. So the minimum page measurement of *R* (i.e. $c_{\min} = \min_{j \in [1,n]} (\| R'_j \|)$ in section 3) is not too small and

$\delta \geq 0.4$	*maxAlgo*	*avgAlgo*	*minAlgo*
	Threshold=2.999	*Threshold=1.434*	*Threshold=0.999*
No. of Filtered Pages	6496	4704	3808
No. of Filtered Noise Pages	2968	2912	2408
No. of Filtered Suspected Pages	1624	1512	1176
NPFR	**1.00**	**0.98**	**0.81**
NPFP	0.46	0.62	0.63
SPFP	0.25	0.32	0.31
EFP	**0.71**	**0.94**	**0.94**
$\delta \leq 0.3$	*maxAlgo*	*avgAlgo*	*minAlgo*
	Threshold=2.999	*Threshold=1.434*	*Threshold=0.999*
No. of Filtered Pages	6608	5096	4648
No. of Filtered Noise Pages	2968	2912	2912
No. of Filtered Suspected Pages	1624	1512	1344
NPFR	**1.00**	**0.98**	**0.98**
NPFP	0.45	0.57	0.63
SPFP	0.25	0.30	0.29
EFP	**0.70**	**0.87**	**0.92**

Table 1: Numerical results for three algorithms *maxAlgo*, *avgAlgo* and *minAlgo*

No.	URL	Title
1	www.corporate-ir.net	*CCBN: Corporate Communications Broadcast Network*
2	aero-news.net/news/ticker.htm	*AERO-NEWS Network: Aviation News Ticker*
3	www.biochrom.co.uk/biochrom.htm	*Biochrom Ltd manufacturer of Amino Acid Analysers …*
4	www.warnerinstruments.com	*Warner Instrument Corporation*
5	www.theweathernetwork.com/cities/can/Tillsonburg_ON.htm	*The Weather Network - Weather Forecast - Tillsonburg*
6	www.hugo-sachs.de	*Hugo Sachs Elektronik*
7	www.nrc.ca/inms/time/cesium.shtml	*NRC Time Services: Web Clock*
8	www.unionbio.com	*Welcome to Union Biometrica*
9	www.mitoscan.com	*MitoScan rapid mitochondria*
10	htmlgear.lycos.com/specs/guest.html	*Html Gear - Gear Specification - Guest Gear*

Table 2: Ten arbitrary noise pages

No.	URL	Title
11	search.harvard.edu:8765	*Search Harvard University*
12	www.harvard.edu/listing	*Index to Harvard University web sites*
13	www.harvard.edu/about	*About Harvard University*
14	www.harvard.edu/academics	*Harvard University: Academic programs*
15	www.harvard.edu/admissions	*Harvard University: Admissions offices*
16	www.haa.harvard.edu	*An Online Community for Harvard University Alumni*
17	www.workingatharvard.org/em-main.html	*Harvard University Office of Human Resources,Employment*
18	www.news.harvard.edu	*Harvard University News Office*
19	www.athletics.harvard.edu/admstaff.html	*Harvard University Athletics: Administrative/Coaching Staff*
20	www.athletics.harvard.edu/vsports.html	*Harvard University Athletics: Varsity Sports*

Table 3: Ten arbitrary topic-related pages

minAlgo algorithm is suitable for small δ value. However, according to our experimental experience, if the linkage distribution within the root set is not even, the minimum page measurement of R may be too small and many noise pages cannot be eliminated. In that case, only *avgAlgo* algorithm is suitable. Therefore, we suggest and adopt *avgAlgo* algorithm as a suitable algorithm for eliminating noise pages in most cases, and in practical computation, the value of parameter δ could be chosen as 0.5.

It has been mentioned in section 3 that our algorithm enables the topic-related pages to capture main linkage information among the pages. That means topic-related (term-related) pages should keep main linkage information, while the noise pages should keep less linkage information. In other words, when the value of parameter δ changes from large to small, the decrease of page measurements, which are defined in (7), of noise pages would be much greater than that of topic-related

Page No.	1	2	3	4	5	6	7	8	9	10
$\delta = 0.5$	0.937	0.957	0.937	0.937	0.957	0.937	0.957	0.937	0.937	0.957
$\delta = 0.3$	0.000	0.006	0.000	0.000	0.006	0.000	0.006	0.000	0.000	0.006
Decrease rate	100%	99%	100%	100%	99%	100%	99%	100%	100%	99%
Avg. decrease rate	99.6%									

Table 4: Page measurement changes of noise pages with different values of parameter δ

Page No.	11	12	13	14	15	16	17	18	19	20
$\delta = 0.5$	1.844	4.967	2.568	3.203	2.518	1.844	1.844	3.388	1.186	1.630
$\delta = 0.3$	1.590	3.534	1.796	2.127	1.112	1.590	1.590	3.212	1.014	1.153
Decrease rate	14%	29%	30%	34%	56%	14%	14%	5%	15%	29%
Avg. decrease rate	24%									

Table 5: Page measurement changes of topic-related pages with different values of parameter δ

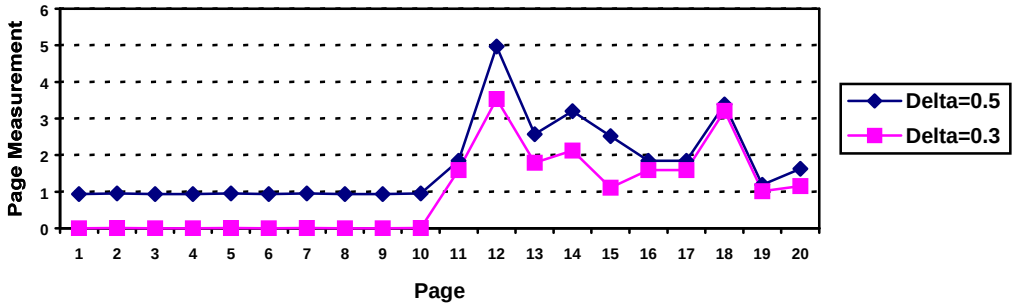

Figure 3: Page measurement change trends for 20 arbitrary selected pages with different values of parameter δ

pages. We arbitrarily choose ten noise pages and ten topic-related pages to see their page measurement changes with the changes of δ value. The ten noise pages and ten topic-related pages are listed in table 2 and table 3 respectively.

Table 4 and figure 3 (pages 1-10) show the page measurement changes of noise pages with the changes of δ value. Table 5 and figure 3 (pages 11-20) show the page measurement changes of topic-related pages with the changes of δ value. It is very clear that when the value of δ changes from 0.5 to 0.3, the page measurements of noise pages decrease at least 99% and average decrease rate is 99.6%. This indicates that noise pages do not capture main linkage information among pages. On the other hand, however, for the same situations, the page measurements of topic-related pages do not decrease too much (at most 56%, at least 5% and average decrease rate is 24%). It suggests that topic related pages capture main linkage information. These numerical results are coincident with the above analysis, i.e. our algorithm enables the topic-related pages to capture main linkage information, while the noise pages not to.

Finally, we apply HITS algorithm to the base set of pages B in which noise pages are not eliminated by our algorithm. On the other hand, we use our algorithm (*avgAlgo* algorithm with δ=0.5) to eliminate noise pages from B and get a new base set B'. We then apply HITS to this new base set B'. The top five authorities and hubs for each situation are listed in table 6 and 7 respectively.

It is indicated from these two tables that before noise pages are eliminated from the base set, HITS produces three authorities (i.e. *http://www.corporate-ir.net*, *http://www.biochrom.co.uk/biochrom.htm* and *http://highwire.stanford.edu*) that have no relationships with the term "*Harvard*" (table 6). After noise pages are eliminated by our algorithm, HITS algorithm produces satisfactory results (table 7), i.e. every produced authority and hub is topic-related. These experiment numerical results indicate that our algorithm really improves base set quality and can be used to construct good quality Web communities.

5 Related Work and Comparison

In Kleinberg's HITS algorithm (Kleinberg 1999), the base set of pages is derived directly from the root set by simply adding pages that point to or are pointed to by the pages in the root set. This will, as we indicated previously, bring many topic-unrelated pages into the base set and finally cause the topic drift problem in many cases. Although Bharat et al. (Bharat and Henzinger 1998) improved HITS algorithm, they only concerned how to reduce the influence of noise pages in the community construction, rather than eliminating these noise pages. In our algorithms, base set of pages is selectively constructed by eliminating (filtering) noise pages. This approach has its intuitive meaning that if a page has fewer links to the pages in root set, it is most likely to be unrelated to the query topic. Therefore, the pages in the base set of pages should have more links to the pages in the root set. However, how many links to the

Top Five Authorities		
Authority value	URL	Title
0.735	www.harvard.edu	*Welcome to Harvard University*
0.285	www.fas.harvard.edu	*Faculty of Arts and Sciences, Harvard University*
0.207	www.corporate-ir.net	*CCBN: Corporate Communications Broadcast Network*
0.190	www.biochrom.co.uk/biochrom.htm	*Biochrom Ltd manufacturer of Amino Acid Analysers …*
0.151	highwire.stanford.edu	*HighWire Press*
Top Five Hubs		
Hub value	URL	Title
0.235	www.fas.harvard.edu	*Faculty of Arts and Sciences, Harvard University*
0.226	post.economics.harvard.edu/info/links.html	*Harvard Economics Links Page*
0.218	www.physics.harvard.edu	*Harvard University Department of Physics*
0.206	www.harvard.edu/academics	*Harvard University: Academic programs*
0.192	www.harvard.edu/listing	*Index to Harvard University web sites*

Table 6: Top five authorities and hubs *before* noise pages being eliminated

Top Five Authorities		
Authority value	URL	Title
0.788	www.harvard.edu	*Welcome to Harvard University*
0.283	www.fas.harvard.edu	*Faculty of Arts and Sciences, Harvard University*
0.213	www.economics.harvard.edu	*Harvard University Department of Economics*
0.191	www.gsas.harvard.edu	*Graduate School of Arts & Science, Harvard University*
0.143	www.law.harvard.edu	*HLS: The Harvard Law School Home Page*
Top Five Hubs		
Hub value	URL	Title
0.244	post.economics.harvard.edu/info/links.html	*Harvard Economics Links Page*
0.238	www.fas.harvard.edu	*Faculty of Arts and Sciences, Harvard University*
0.196	post.economics.harvard.edu/people	*Harvard Economics Directories of Faculty, Staff & Students*
0.195	www.fas.harvard.edu/about	*About Harvard University Faculty of Arts & Sciences*
0.195	www.physics.harvard.edu	*Harvard University Department of Physics*

Table 7: Top five authorities and hubs *after* noise pages being eliminated

root set pages is a reasonable threshold for a page to be included in the base set of pages? In other words, what is the numerical threshold for eliminating noise pages? In our algorithm, this threshold is numerically defined (i.e. *avgAlgo* algorithm with δ=0.5), and the closeness of a page to the root set is also numerically measured.

There are other methods that are based on the page content analysis for eliminating noise pages or reducing the influence of noise pages. For example, in (Bharat and Henzinger 1998), the similarity between a page and the query is used to determine whether this page is a noise page or not. For a given threshold, if the similarity between a page and the query is below this threshold, this page will be eliminated from the base set. But this similarity is determined by the page contents and refers to building the source of *IDF* weights (IDF_i is an estimate of the Inverse Document Frequency of term *i* on the WWW) for the Web. (Chakrabarti, Dom, Gibson, Kleinberg, Raghavan, and Rajagopalan 1998) takes a similar way to get the similarity between a page and the query by analysing the page contents around the link. In both (Chakrabarti, Dom, Gibson, Kleinberg, Raghavan, and Rajagopalan 1998) and (Bharat and Henzinger 1998), page similarity is also used to change the linkage weight between two pages, and in turn to reduce the influence of noise pages. Since HTML page tags on the Web contain few semantics, it is usually hard to exactly analyse the content of the Web pages, and these content analysis based methods have limitations in their applications.

Our algorithm purely exploits the linkage information, which is easy to be extracted, between the Web pages to analyse the relevance between the pages. Based on the SVD of the linkage matrix, our algorithm enables the topic-related pages to capture the main linkage information and the most likely noise pages to capture the minor linkage information. The algorithm reveals the relationships between the pages at a deeper level, and makes it possible to separate and eliminate most likely noise pages from the original base set of pages. By combining our improved base set of pages with community construction algorithms (e.g. HITS, those in (Bharat and Henzinger 1998) and (Chakrabarti, Dom, Gibson, Kleinberg, Raghavan, and Rajagopalan 1998)), good quality Web communities can be constructed.

6 Conclusions

In this paper, we propose a noise page elimination algorithm (NPEA) to eliminate noise pages from the base set of Web pages and improve the quality of base set, which in turn makes it possible to construct a good quality Web page community. Based on the basic linkage information between the pages, the algorithm reveals the relationships among the concerned pages at a deeper level and numerically defines the threshold for eliminating noise pages. The numerical experiment results show the effectiveness and feasibility of the algorithm. The ideas in the algorithm can also be used in other hyperlink analysis. Further more, this algorithm could also be used solely to

filter unnecessary Web pages and reduce the management cost and burden of Web-based data management systems.

Since relationships among the pages in our algorithm neglect minor influence factors (sparse linkages) and capture the dense linkages, they can also be used to analyse the similarity between the pages and derive clustering algorithms for the Web pages. The similarity between Web pages can be further used for Web page retrieval on the WWW. These will be our further research directions.

7 References

AltaVista, *http://www.altavista.com/*.

BHARAT, K. and HENZINGER, M. (1998): Improved Algorithms for Topic Distillation in a Hyperlinked Environment. *Proc. the 21st International ACM Conference of Research and Development in Information Retrieval (SIGIR98)*, 104-111.

CHAKRABARTI, S., DOM, B., GIBSON, D., KLEINBERG, J., RAGHAVAN, P. and RAJAGOPALAN, S. (1998): Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. *Proc. the 7th International World Wide Web Conference*, 65-74.

DATTA, B.N. (1995): *Numerical Linear Algebra and Application*. Brooks/Cole Publishing Company.

DEAN, J. and HENZINGER, M. (1999): Finding Related Pages in the World Wide Web. *Proc. the 8th International World Wide Web Conference*, 389-401.

DEERWESTER, S., DUMAIS, S.T., FURNAS, G.W., LANDAUER, T.K. and HARSHMAN, R. (1990): Indexing by Latent Semantic Analysis. *J. Amer. Soc. Info. Sci*., 41(6): 391-407.

GIBSON, D., KLEINBERG, J. and RAGHAVAN, P. (1998): Inferring Web Communities from Link Topology. *Proc. the 9th ACM Conference on Hypertext and Hypermedia (HyperText98)*, 225-234.

GOLUB, G.H. and VAN LOAN, C.F. (1993): *Matrix Computations, second edition*. The Johns Hopkins University Press.

HOU, J., ZHANG, Y., CAO, J., LAI, W. and ROSS, D. (2001): Visual Support for Text Information Retrieval Based on Linear Algebra. *Journal of Applied Systems Studies*, Cambridge Scientific Publishers (to appear).

HOU, J., ZHANG, Y., CAO, J. and LAI, W. (2000): Visual Support for Text Information Retrieval Based on Matrix's Singular Value Decomposition. *Proc. of the 1st International Conference on Web Information Systems Engineering (WISE'00)*, Hong Kong, China, Vol.1 (Main Program): 333-340.

KLEINBERG, J. (1999): Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM* 46(1999).

Yahoo!, *http://www.yahoo.com/*.

8 Appendix

Noise Page Elimination Algorithm (NPEA)

NPEA ($G(R)$, $G(B)$, δ)

Input:

***G(R)*:** $G(R)=(R,E_R)$ is a directed graph of root set pages with nodes being pages and edges being links between pages.

***G(B)*:** $G(B)=(B,E_B)$ is a directed graph of base set pages with nodes being pages and edges being links between pages.

***δ* :** threshold for selecting matrix approximation parameter k.

Output:

***G'(B)*:** a new directed graph of base set pages without noise pages.

Begin

Get the number of pages in B, m = size(B);
Get the number of pages in R, n = size(R);
Construct linkage matrix between pages in R

$S = (s_{ij})_{n\times n}$;

Construct linkage matrix between B-R and R

$A = (a_{ij})_{(m-n)\times n}$;

Compute the SVD of S and its singular values

$S_{n\times n} = W_{n\times n}\Omega_{n\times n}X^T_{n\times n}$;

$\omega_1 \geq \omega_2 \geq ... \geq \omega_t > \omega_{t+1} = ... = \omega_n = 0$;

Compute the SVD of A and its singular values

$A_{(m-n)\times n} = U_{(m-n)\times(m-n)}\Sigma_{(m-n)\times n}V^T_{n\times n}$;

$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > \sigma_{r+1} = ... = \sigma_n = 0$;

Choose parameter k such that

$(\sigma_k - \sigma_{k+1}) / \sigma_k \geq \delta$;

Compute coordinate vectors R_i (i = 1, 2, …, m-n) for each page in B-R according to (3);

Compute coordinate vectors R'_i (i = 1, 2, …, n) for each page in R according to (4);

Compute the projection vectors PR_i (i = 1, 2, …, m-n) according to (6);

Compute the representative measurement of R

$c = \sum_{j=1}^{n} \| R'_j \| \Big/ n$;

if $\| PR_i \| < c$ (i = 1, 2, …, m-n) **then**

Begin

Eliminate page i from B

$B = B$ - *page* i;

For every page $p \in B$, eliminate links with page i from E_B

$E_B = E_B$ - *(page* $i \rightarrow p$*)* - *(*$p \rightarrow$ *page* i*);*

End

return $G'(B)=(B,E_B)$;

End