Available online at www.sciencedirect.com
SCIENCE DIRECT®
Decision Support Systems 41 (2006) 574–591
Decision Support Systems
www.elsevier.com/locate/dsw

Mining web navigations for intelligence

Harris Wu [a,*], Michael Gordon [b], Kurtis DeMaagd [b], Weiguo Fan [c]

[a] *Old Dominion University, Norfolk, VA 23529, United States*
[b] *University of Michigan Business School, 701 Tappan Street Ann Arbor, MI 48103, United States*
[c] *Virginia Tech, Blacksburg, VA 24061, United States*

Available online 3 September 2004

Abstract

The Internet is one of the fastest growing areas of intelligence gathering. We present a statistical approach, called principal clusters analysis, for analyzing millions of user navigations on the Web. This technique identifies prominent navigation clusters on different topics. Furthermore, it can determine information items that are useful starting points to explore a topic, as well as key documents to explore the topic in greater detail. Trends can be detected by observing navigation prominence over time. We apply this technique on a large popular website. The results show promise in web intelligence mining.

Keywords: Principal clusters analysis; Intelligence; Mining; Trend analysis; Navigation analysis; Information overload; Web community

1. Introduction

The Internet is one of the fastest growing areas of intelligence gathering. On Nov. 25, 2002, President Bush signed the Homeland Security Bill into law, which adds greater Internet surveillance power to authorized agencies. However, there is too much information on the Internet and too many activities to monitor. Such information overload presents a formidable challenge to intelligence workers striving to analyze this substantial collection of data without being overwhelmed by the effort required. This study develops an approach that analyzes web navigations to help intelligence workers. By capturing and analyzing web navigations, we identify prominent topics and the most important documents on these topics. Once identified, these can be used by intelligence workers to take proper actions. For example, emerging topics can be used to find new security threats. Key documents on a certain topic may then lead to identification of key personnel in terrorism groups.

Much research has been done on link analysis in the last decade, in both the fields of information retrieval and network analysis. On the Internet, hyperlinks between web documents are easy to capture because they are contained in static documents which can be retrieved by web crawlers. Hyperlink data, however,

* Corresponding author.
E-mail addresses: hwu@odu.edu (H. Wu), mdgordon@umich.edu (M. Gordon), demaagd@umich.du (K. DeMaagd), wfan@vt.edu (W. Fan).

0167-9236/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2004.06.011

lack information about users' dynamic behavior. Navigation analysis has the potential to study such behavior.

In this paper, we present an approach to analyze web navigations. In particular, we identify prominent topics or clusters of user navigations and important documents within these topics. We empirically evaluate this approach on a large, heavily visited website. In Section 2, we review previous research on web navigations. In Section 3, we describe our research site and data collection. In Section 4, we present the statistical technique we have developed, which identifies the structure in the navigation data represented by weighted link matrices. In Section 5, we present our results so far on our testing site. These results indicate that our approach is successful in finding topics of user activity and key documents related to these topics, which will support intelligence. Our approach successfully handles very large web-based document collections and user bases. Section 6 concludes the paper with a discussion of potential applications of our research and future research directions.

2. Literature review

2.1. Navigation studies

Much research has focused on the link structure between documents. Bibliometrics utilizes quantitative analysis and statistics to describe patterns of publication within a given field or body of literature. Citation indexes such as the Social Science Citation Index have been used to suggest prominent literature and individuals in a field. In recent years, a new growth area in bibliometrics has been on analyzing web hyperlinks. For example, Google uses hyperlink structure to determine the ranking of search results.

While hyperlinks are static links contained in documents, web navigations are dynamic, being based on people's online behavior. In addition to being useful for studying the structure of a document collection, navigation data can be used to study people's online activities. Navigation patterns can be studied for either an individual or a group. Navigation patterns can also be studied for certain periods of time, over which trends can be identified.

From an analysis perspective, both hyperlinks and navigations can be represented by link matrices with outgoing/start documents as rows and incoming/destination document as columns. In hyperlink analysis, the values in the matrix are either 0 or 1, as two documents are either linked or unlinked. In navigation analysis, links between documents are weighted by the frequency of navigation. In this sense, navigation data are richer than hyperlinks.

Some early research on web navigations were concerned with an individual user's browsing pattern and with browser usability. Catledge and Pitkow [5] characterized users' different browsing strategies. Cockburn and Jones [8] analyzed inadequacies of browsers with usability studies. Tauscher and Greenberg [40] reported on users' revisitation patterns of web pages. Choo et al. [7] studied the information-seeking behavior of managers and IT specialists. In their findings, people tend to browse in very small clusters of related pages, and generate only short sequences of repeated URL paths.

Web navigations have also been studied from a website usability perspective. Research has used navigation data to perform task analysis on a website, such as to study the time to complete a task [16]. Click-stream analysis tools are used to study return-on-investment of advertisements or web pages (e.g., Ref. [1]). Research has also tried to profile users based on their navigation patterns (e.g., Refs. [18,29]).

A large body of web navigation research is related to information retrieval. One simple aspect of Web navigation, page hits or the *frequency of document access*, has been used to rate and recommend web documents. Garofalakis et al. [19] studied website optimization using page popularity. Joachims [23] analyzed which URLs returned by search engines are actually clicked through by users, and found that a search engine's performance could be improved by ranking more frequently followed URLs higher. The *time of document access* is also informative, as documents accessed by users in the same time period can be viewed as related. Perkowitz and Etzioni [33] clustered documents that were accessed by the same users within a 1-day period. Su et al. [39] discussed clustering algorithms for pages based on the temporal proximity of users' access to pages. The *path of document access* has often been studied in the context of adaptive hypertext systems. Some systems periodi-

cally reorganize or dynamically update site structure according to user needs learned from users' traversals (e.g., Ref. [38]). Chen et al. [6] predicted the types of documents a user will access in a structured document collection. Zhu et al. [45] used Markov models to predict individual users' destinations based on their recent navigation history.

Other research extracts association rules, sequential patterns, clusters, path expressions or other structures from user navigation using a data mining approach (e.g., Refs. [6,22,35–37,44]). However, the output of these mining studies is often not in a form suitable for direct human consumption [9]. The mining results are often so large that they need to be stored in a data warehouse for future queries and analysis (e.g., Ref. [25]). Visualization tools may help interpret the results (e.g., Refs. [11,34]), although construction of an effective visualization is a challenging task itself.

The research above is chiefly targeted to help software developers, web users and website owners. However, little navigation research has been devoted to intelligence workers such as the analysts in security or intelligence agencies. So far, no research has reported mining user navigation from a sizable environment to gather intelligence directly understandable by human analysts. Part of the reason may be that user navigation is not always easy to capture in detail, and once captured, the data volume may be too huge to analyze and distill. Next, we will talk about data collection and mining in more detail.

2.2. Navigation data collection mechanisms

Some previous research (e.g., Refs. [7,40]) has used client-side logging, where clients are instrumented with special software so that all browsing behavior is captured. The advantage of client-side logging is that literally everything can be recorded, from low-level events such as keystrokes and mouse clicks to higher-level events such as page requests, all of which can be valuable information. Furthermore, there are some client-side logging tools readily available (e.g. Refs. [15,41]). However, there are several drawbacks to client-side logging. Special software must be installed on the client computers, which users may be unwilling or unable to do. The software only works for specific operating systems or specific browsers. There also needs to be some mechanisms to collect the logged data. In short, client-side logging has limited use in intelligence gathering.

Navigations within a website can be observed using web server logs. Each log record typically contains the timestamp, the URL visited and the originating IP address. Hits from an IP address within a certain time frame are assumed to come from a single user session for a certain task; thus, visitors' navigation sequences are reconstructed from the logs. Web server logs are frequently used to study web navigations because nearly every web server can automatically log page requests and these logs are conveniently available. Furthermore, there are many web server log analysis tools readily available, with over 100 commercial and free tools currently on the market (http://www.uu.se/Software/Analyzers). Interpreting the actions of an individual user from web server logs, however, has many problems as pointed out by Etgen and Cantor [16], Davison [12] and Hong et al. [21]. Caches such as client browser caches and Intranet or ISP caches can intercept requests for web pages. If the requested page is in the cache, then the request will never reach the web server and is thus not logged. Multiple people sharing the same IP address, a general practice by ISPs such as American Online, makes it difficult to distinguish who is requesting what pages. Dynamically assigned IP addresses, where a computer's IP address changes every time it connects to the Internet, can also make it difficult to determine what an individual user is doing since IP addresses are often used as identifiers. While researchers have found novel ways of extracting useful user path data from server logs on a statistical level [31], the exact paths of individual users still remain difficult to extract. Despite these issues, many recent studies use web logs (e.g., Ref. [45]).

Proxies are widely used in corporations, universities or other organizations, typically for security and ease of administration. A proxy server lies between browsers and the Internet, and captures all browser page requests. If it can, it returns pages from its cache; otherwise, it requests the page from the Internet. Most proxy servers can generate proxy logs. Unlike web logs, a proxy can capture user navigations on the whole Internet instead of within a web site. Proxy logs are also more accurate than web logs in matching users and webpage requests. A proxy does not involve deploy-

ment on either the client side or server side, and is thus transparent to both Internet users and web administrators. There are many proxy log analysis tools available, some of which can analyze both web logs and proxy logs (e.g., Flowerfire:http://www.sawmill.net). For a collocated group of users with shared Internet access, or organizational users with central firewalls, a proxy server is a good way to capture user navigations. For example, Wang et al. [42] used proxy logs from a proxy server at Microsoft. The proxy approach can be extended to distributed users. Hong et al. [21] developed Webquilt, a URL-based proxy for remote web usability testing, which allows distributed users to voluntarily go through a website-based proxy, which redirects and captures users' browsing requests. A proxy can work well for a group of cooperative users, but not for intelligence gathering among general Internet users. However, routers or packet snoopers on the network can potentially monitor the navigation behavior of any users using the network (e.g., Ref. [14]). The router or snooper based approach is most promising for intelligence gathering on the open Internet by authorized agencies.

Another way to capture user navigations within a website is to use server-side scripts. Many websites serve content dynamically using server-side scripts, such as Common Gateway Interface (CGI) scripts, instead of static HTML pages. It is not difficult to add code to these server-side scripts to capture user sessions. Because the content is dynamically served, there are no caching issues. The server-side scripts can accurately tell which user is requesting what. In fact, many advertisement links on the web capture user click sequences using server-side scripts, and website owners get paid based on how many people have clicked on these links. Moreover, since each client browser has a separate interaction with the server, the server-side scripts can track simultaneous browsing sessions on different tasks by the same user, which cannot be extracted from web logs or proxy logs. In our research, we use embedded coding in server-side scripts to capture user navigations. The techniques we use to analyze user navigations, however, apply to any means of capture.

2.3. Navigation mining techniques

A human analyst only has the capacity to process a small number of items within a large amount of navigation data from a large document collection. Thus, to understand users' behavior, it is important either to produce a few key representative items, such as documents or navigation paths, or a few aggregate items, such as topics.

Tauscher and Greenberg [40] used a pattern detection approach [10] to identify the longest repeated sequences (LRS) in user navigation. Papadimitriou et al. [30] extended the approach with a weighted specificity rule, which weights longer sequences higher. These longest repeated sequences tend to cover highly traversed documents. However, those pages that users go to directly may not be covered by longest repeated sequences; thus, the LRS may give a distorted view of user behavior.

Document quality can be determined using page request statistics and link analysis techniques. The HITS algorithm [26] can identify important documents including those that refer to many popular pages (hubs) or those referred to by many popular pages (authorities). Kleinberg's hub/authority approach has been applied to user navigation [42]. However, these highly ranked documents can only be identified relative to a given query or a set of seed documents on a certain topic.

Much research has been devoted to finding clusters from user navigation. The clusters may be clusters of documents, navigation paths, or user sessions. For example, Pirolli et al. [32] used a spreading activation model to identify the locality of documents starting from any source document in a website. Perkowitz and Etzioni [33] and Su et al. [39] clustered web pages based on similarities defined by temporal proximity in logs. To generate k clusters out of navigation paths, Shahabi et al. [35] used a cosine measure between paths to construct a similarity matrix, and then used k-mean clustering on the similarity matrix. Joshi et al. [25] used a similar approach to produce k clusters of user sessions. Note that clusters that result from data mining may not be directly understandable by human analysts. However, analysts can use them with some a priori knowledge of the data, e.g., in finding documents similar to an existing document.

Few approaches can provide intelligence analysts a summary of user navigation in a large environment. One approach may be to identify some important documents first using either heuristics or statistics, and then to use these documents as sources to identify

localities that together represent most of the user activity. Another approach may be to identify clusters first, and then distill topics or important documents from these clusters. However, there is no unified approach that can identify topics or clusters that account for most user navigations and the key documents within these topics. We try to address this gap in our paper.

3. Data collection

3.1. Research site

We collected data for our research from http://everything2.org, a large, open, web-based community that encourages users to create, coauthor and discuss documents on a, hypothetically, unlimited set of subjects. The website is developed and maintained by the open source community. All documents within the website are contributed by individual users. As of our last data collection, http://everything2.org had about 100,000 active named users who had logged on in the three previous months. There are about 500,000 documents on a variety of topics. The website receives over 50,000 page views per day. It is worth mentioning that http://everything2.org is a true community in that there are intensive interactions among users in various forms, such as coauthoring, commenting on each other's work, online chat and peer mentoring among the members. Since the community is open to everyone, it is the type of community that may contain crime-related documents and attract users who author or read them. However, due to the scale of the community it is difficult to censor contributed content or monitor individual users' activities. We used this site to help answer the types of questions of interest to intelligence analysts studying online communities. How can one obtain a high-level view of user activities? How can key documents involved in these user activities be identified? Before we try to answer these questions, we first go into more detail on the research site and how user navigation is captured.

Each document in http://everything2.org can be displayed as a web page and has a title that is unique in the website. To explore a certain subject, the easiest way is to type the subject into a search form and perform a title search. If there is a document whose title matches exactly what is typed into the search form, this document will be returned by the title search and displayed in the browser. Otherwise, the search will return a list of documents whose titles are close matches to the subject, and the user can choose to explore any of them. If the user is logged in, below the close matches there is a form to allow the user to create a new document titled as the search input. Besides the title search, the website has more advanced search mechanisms such as full-text search with various search options. The search mechanisms generate navigation links linking a search result with the document the user was viewing when conducting the search.

Another way to navigate http://everything2.org is to follow links contained in the documents. http://everything2.org is a self-contained website, meaning that most links within user-contributed documents refer to documents within the website. Only administrative users can create links that point outside the website. To accommodate users who are not familiar with HTML, links within http://everything2.org are created using a bracket notation. For example, "I enjoy [classic rock] more than. . ." will contain a link that refers to "classic rock".

All types of links are dynamically translated to URLs when the page is served by the system. In the above example, when a user clicks on the "classic rock" link, a server-side script will be invoked to retrieve the document titled "classic rock". If there is no such document, then the server-script will return a list of pages with related titles just as if the user had done a title search on "classic rock". The ease of creating keyword links results in an abundance of them in http://everything2.org.

Server-side scripts on http://everything2.org capture a user's navigation sequence among the documents, also called nodes. Each document has a unique numeric identifier, or node_id. For example, the document "classic rock" has node_id=62311. When "classic rock" is delivered to a user's browser, the links in the document are translated to URLs containing "index.pl?lastnode_id=62311&. . ." Similarly, the search form now contains a hidden input, "<input type="hidden" name="lastnode_id" value="62311">." In other words, the user's session contains a lastnode_id variable that keeps track of the current document.

We use the term *step* to denote navigating from document A to document B, by means of following a link or conducting a search. The *weight* of step (A and B) is defined as the number of times that users have navigated from document A to document B. The server-side scripts record user navigations in a back-end database table using records of the form (from_doc, to_doc, weight). A new record with weight=1 is created the first time a step is navigated, and incremented thereafter.

Note that the server-side scripts can capture individual user navigations in great detail, such as in the format of (user_id, timestamp, from_doc, to_doc). However, due to privacy issues and constraints of server capacity, in our research we only collected data covering users' collective navigations, not for individual users. We performed data collections for two extended, contiguous time periods, periods 1 and 2 (exact timeframes are omitted for confidentiality).

3.2. Characteristics of navigation data

Table 1 lists some descriptive statistics for the navigation data collected from http://everything2.org. The collected data can be represented in a navigation matrix **A** with rows as starting documents and columns as ending documents of navigation steps. For period 1, the navigation matrix has dimension 336,752×326,744. The number of nonzero entries in the matrix, or the number of steps, is 4,881,298. The value of a nonzero entry A_{ij}, is the weight of the step (i, j). The sum of all nonzero values in the matrix is the total number of navigations in period 1, i.e., 13,415,981.

The mean value of weight, or the average number of times that a unique navigation step (A and B) is followed, is 2.7. The standard deviation of weight is 8.5. Fig. 1 shows the weight distribution among all steps. It is clear that weight is not normally distributed. We hypothesize that the step weights follow a power law, similar to hyperlinks and other scale-free networks [2]. Using a log-log linear regression, we obtain an estimate of the distribution function as a power function: Probability(weight=x)=0.242...$x^{-0.274}$. The R^2 is 0.903, which indicates a very good fit. The degree exponent is −2.274, with significance at a 10^{-6} level. In comparison, Barabasi [2] studied hyperlink structures on the web, using outgoing links from and incoming links to web pages. Barabasi's study found that the in-links had a degree exponent of −2.1 and out-links had a degree exponent of −2.5. To further verify our choice of power function model, we used the curve estimation feature in SPSS™. Out of 11 available curve functions in SPSS™, our power function comes out as the best curve estimate. As far as we know, our research is the first to show that navigations have a power-law distribution similar to hyperlinks. The power-law weight distribution violates the multi-normality assumption of many existing multivariate analysis techniques such as Factor Analysis and various classification methods [24].

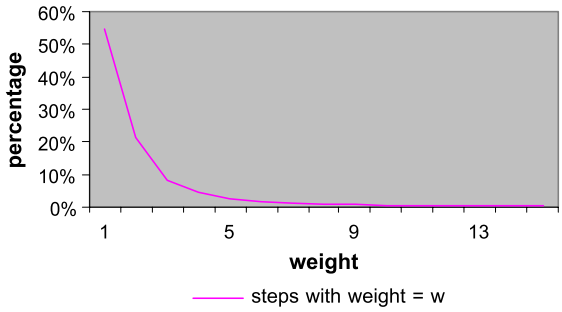


Fig. 1. Distribution of weight of steps.

Table 1
Descriptive statistics for navigation data from periods 1 and 2

Period	No. of distinct starting documents	No. of distinct ending documents	No. of steps	No. of navigations, or total weight
1	336,752	326,744	4,881,298	13,415,981
2	340,363	344,627	4,787,504	12,152,625

If we were to load the matrix **A** in popular statistical packages that store matrices in a two-dimensional array format, the matrix would take over 220 GB of storage with 2 bytes for each entry. Fortunately, the navigation matrix is very sparse. For period 1 only, about one out of 25,000 entries in the navigation matrix have nonzero values. Using the Compressed Column Storage (CCS) or Compressed Row Storage (CRS) format for sparse matrices [3], the matrix can be stored in 40 MB. In short, navigation data is huge and sparse in terms of matrix representation, and may call for special analysis techniques.

Out of the 4.9 million steps in period 1, 55% have only been navigated once. One way to reduce the amount of data is to study only frequently navigated paths (e.g., Ref. [33]); that is, to leave out all entries in the navigation matrix below a certain threshold, say 50 or 100. However, it is difficult to determine the threshold value or to justify the threshold choice. Furthermore, weak links (links of low weight) are known to be critical in connecting different components in networks [20]. It is desirable to take these less navigated paths into consideration.

4. Data analysis

4.1. Principal clusters analysis

As we have seen, the navigation data can be huge. To make sense out of such a large amount of data, we have developed a method called *principal clusters analysis* [43] that identifies major topics from user navigation activities and the most useful documents on each topic. Our notion of principal clusters analysis relates to principal components analysis (PCA) and cluster analysis, two popular techniques in multivariate analysis. PCA reduces the number of variables used to describe a set of data by combining correlated variables. Cluster analysis groups objects similar to each other into groups called clusters.

As does principal component analysis, principal clusters analysis provides a compact description of data. Similar to cluster analysis, it produces a number of groups from a large amount of data. In essence, principal clusters analysis is a data reduction procedure using a truncated version of singular value decomposition (SVD). We apply principal clusters analysis to our navigation matrix to find patterns from the structure of the underlying data.

Step 1 of principal clusters analysis is to decompose the navigation matrix into the underlying dimensions, or topics, using SVD. Recall that the navigation matrix depicts the *source-destination* links in a navigation graph among document nodes. After decomposing this matrix into its singular values, we obtain a matrix (of singular values) that corresponds to the underlying dimensions of the original navigation matrix, plus two other matrices—one that maps the "source" of popular-step links to these underlying dimensions; and another that maps the "destination" of these links to the same underlying dimensions. Mathematically, the original navigation matrix **A** is decomposed as follows:

$$\mathbf{A} = \mathbf{U} \times \mathbf{S} \times \mathbf{V}^{\mathrm{T}}$$

where **U** and **V** are the *source-underlying_dimension* and the *destination-underlying_dimension* matrices, respectively; and **S** is a diagonal matrix of singular values.

Step 2 selects the n largest topics that account for most of the activity in the original navigation matrix. Mathematically, **A** is reduced to the n largest singular values from **S**) along with the corresponding first n column vectors in **U** and **V** (i.e., the top n left singular vectors and top n right singular vectors, respectively). **A** is now approximated by the n largest topics:

$$\mathbf{A}_n \sim \mathbf{U}_n \times \mathbf{S}_n \times \mathbf{V}_n^{\mathrm{T}}$$

$\mathbf{A}_n$ is also called the rank-n truncated SVD of **A**. Among all rank-n matrices **B**, $\mathbf{B}=\mathbf{A}_n$ is the unique minimizer of the Frobenius norm of **A**–**B** (in other words, **B** is a good approximation of **A**):

$$\|\mathbf{A}-\mathbf{B}\|_{\mathrm{F}} = \left(\sum \left(\mathbf{A}-\mathbf{B}\right)_{ij}^{2}\right)^{1/2}.$$

$\mathbf{B}=\mathbf{A}_n$ also minimizes the spectral norm, or the L-2 norm of **A**–**B**:

$$\|\mathbf{A}-\mathbf{B}\|_2 = \max\left\{\lambda^{1/2}, \lambda \text{ is an eigenvalue of } (\mathbf{A}-\mathbf{B})^{\mathrm{T}} \times (\mathbf{A}-\mathbf{B})\right\}.$$

In other words, $\mathbf{A}_n$ is a good approximation of **A**, in terms of both the raw number of navigations and the dimensions of navigation activity. The approximation of **A** by $\mathbf{A}_n$ can be measured by $(\|\mathbf{A}_n\|_{\mathrm{F}}/\|\mathbf{A}\|_{\mathrm{F}})^2$, which is equal to the sum of squares of top n singular values divided by sum of squares of all singular values [24].

Step 3 identifies important start nodes and end nodes in users' navigations for each topic. Mathematically, we "shorten" all the left singular vectors by (a) identifying the k positions in each of the left singular vectors that have the largest absolute values, also called loadings; and then (b) collapsing the left singular vector matrix onto (the union of) these

positions. The right singular vectors are "shortened" similarly to contain only j positions in each vector, j not necessarily=k. **A** is now approximated by

$$\mathbf{A}_{n/j,k} \sim \mathbf{U}_{n|k} \times \mathbf{S}_n \times \mathbf{V}^{\mathrm{T}}_{n|j}$$

In selecting the nodes with the largest loadings, we hope to identify important start nodes and end nodes in users' navigations from the left and right singular vectors. We call important start nodes and end nodes *hubs* and *authorities,* respectively, following Kleinberg [26]. We call the loadings of hubs or authorities hub scores or authority scores. We use the term *principal cluster* to denote a set of documents containing hubs and authorities associated with a top singular value, and navigations among them.

Together, Steps 1–3 convert the navigation matrix to a much simplified representation that identifies prominent topics of navigation activities, important starting nodes for navigation on a topic, and the most useful destinations on a topic.

We use a small matrix to illustrate how the principal clusters analysis works. In Fig. 2, our method can reduce the navigation matrix A (which shows navigations among eight nodes) to two navigation clusters, corresponding to the two largest singular values (74.9 and 40.9) in matrix **S**. The first cluster has node 3 as the top hub and node 4 as the top authority, as indicated by the largest absolute values in the first column vectors in **U** and **V**, respectively. The second cluster has node 1 as the top hub and node 7 as the top authority, as indicated by the second column vectors in **U** and **V**. By only picking the top hubs and the top authorities from the top 2 clusters, we can reduce the 8×8 navigation matrix **A** to $\mathbf{A}_{2|1,1}$, which contains two clusters, each with only one hub and one authority. (Refs. [3,4] and Refs. [1,6]).

4.2. Some related SVD-based techniques

Singular value decomposition has been used to reduce many complex problems. Two recent SVD-

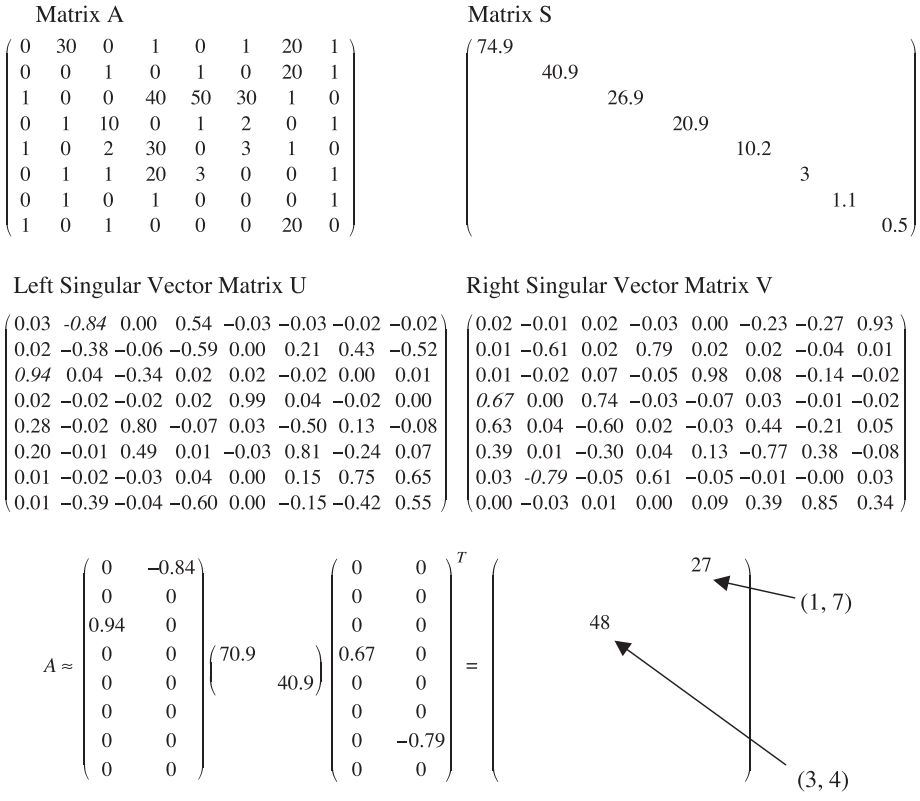


Fig. 2. Matrix decomposition $\mathbf{A}=\mathbf{U}\times\mathbf{S}\times\mathbf{V}^{\mathrm{T}}$ and its approximation.

based techniques for information retrieval are Latent Semantic Indexing [13] and Hub/Authority analysis [26].

Latent Semantic Indexing (LSI) applies SVD to a term-document matrix to reduce dimensionality. Suppose we have a set of ***d*** documents and a set of ***t*** terms. We model each document as a vector ***x*** in the t-dimensional space. The jth coordinate of ***x*** is a number that measures the association of the jth term with respect to the given document, which is generally defined to be 0 if the document does not contain the term, and nonzero otherwise. Now the matrix $\mathbf{X}$ containing these d vectors as columns has dimension $t \times d$. By retaining only the k largest singular values in the SVD of $\mathbf{X}$, we have a rank-k truncated SVD of $\mathbf{X}$: $\mathbf{X}_k = \mathbf{U}_k \times \mathbf{S}_k \times \mathbf{V}_k^{\mathrm{T}}$, with $\mathbf{U}_k$ a $t \times k$ matrix and $\mathbf{V}_k^{\mathrm{T}}$ a $k \times d$ matrix. Each of the d columns of the matrix $\mathbf{S}_k \times \mathbf{V}_k^{\mathrm{T}}$, a $k \times d$ matrix, represents one of the documents. That is, the documents have been projected into the k-dimensional space spanned by the columns of the matrix $\mathbf{U}_k$. Each term no longer occupies a distinct dimension either. Rather, each of the k new dimensions corresponds to a column vector in $\mathbf{U}_k$, which is a weighted sum of terms. These vectors are said to represent the fundamental "concepts" that underlie the collection of documents [28].

Similarly, we try to use the singular vectors associated with top singular values to discover concepts or topics. The claim that LSI can uncover meaningful topics has been analytically investigated by Papadimitriou et al. [30] using a probabilistic model. In their model, topics are a probability distribution over terms. The document collection consists of k different topics hidden from the retrieval algorithm. A document on a given topic τ is generated by a random selection of terms from a probability distribution F_τ over terms. For different topics τ and τ', there is a technical condition enforcing that the distributions F_τ and F'_τ are well separated. Their analytical study shows that when the distributions induced by different topics are sufficiently separated, the k-dimensional subspace produced by LSI yields, with high probability, sharply defined clusters among documents of the k different topics as measured by their cosines. Although Papadimitriou et al.'s study is on Latent Semantic Indexing, it suggests that the SVD in our principal clusters analysis is able to uncover meaningful topics in underlying navigation data.

Hub/authority analysis [26] is a technique for locating high-quality information related to a given search topic on the web. The technique is based on the premise that hyperlinks confer authority; that is, they convey quality endorsements. For a given search topic, a query is first issued to the search engine and the highest ranked documents are returned. A document collection is then constructed using these documents and documents linking to or from these documents. Within the document collection, there are a number of authoritative pages or so-called authorities that have many in-links. In addition, there is a set of hubs, which are pages containing links to many relevant authorities. Hubs and authorities exhibit a mutually reinforcing relationship: a good hub is a page that points to many good authorities; and a good authority is pointed to by many good hubs. Kleinberg uses an iterative algorithm to maintain and update a numerical hub weight and an authority weight for each page in the collection. The documents with largest converged weights are hubs and authorities. In essence, the algorithm is equivalent to finding the left and right singular vectors associated with the largest singular value of the hyperlink adjacency matrix $\mathbf{A}$ for the constructed document collection.

For a broad topic, singular vectors associated with nonlargest singular values in Kleinberg's [26] study may provide additional sets of hubs and authorities, and each set appears to be on a different subtopic. Although not fully understood at an analytical level, the subtopics may be found when the query string to the search engine has different meanings, e.g., "jaguar"; or when the query string arises as a term in the context of multiple communities, e.g., "randomized algorithms"; or when the query string refers to a highly polarized issue, e.g., "abortion".

Although hub/authority analysis is on a constructed document collection related to a given topic, the concept of authority conferral and quality ranking can be generalized to a natural collection of documents on different topics. The hyperlink topology can also be generalized to navigation graphs. As an extension of hub/authority analysis, our SVD-based technique can identify sets of hubs and authorities from user navigation activity. Furthermore, a set of hubs densely linking to the set of authorities associated with the same singular value can be viewed as a "community" structure associated with a topic of general interest on

the web. This basic type of linkage pattern has been found as a recurring and fundamental feature of the web (e.g., Ref. [27]). The "community" structure provides another way to understand web activities.

It is worth noting that much research has been devoted to SVD computation. Large-scale singular value computations have been studied using both sequential and parallel algorithms (e.g., Ref. [3]). For an extremely large amount of data, there are approaches that produce highly efficient approximation algorithms through random sampling (e.g., Ref. [17]). SVD updating, the computation of SVD from incremental changes to the original matrix, has also been studied (e.g., Ref. [4]). Similar to LSI and hub/authority analysis, our SVD-based technique is extremely scalable using these available SVD computation techniques.

5. Results and evaluation

Like intelligence mining itself, our analysis of detecting patterns within user navigation data is highly exploratory in its nature. It is difficult to evaluate our results against different types of results from other analysis techniques. However, the validity of an exploratory analysis can be examined from several different angles [24]: external criteria, internal criteria and replicability. For *external criteria*, we evaluate our results against real world experience. For *internal criteria*, we evaluate how much user activity has been described by the principal clusters that result from our analysis. For *replicability*, we compare results using statistical samples of our data. We also compare results from two different time periods to cross-validate our analysis.

5.1. External criteria

5.1.1. A priori hypotheses

We had very limited knowledge about the overall user activity on http://everything2.org prior to our data analysis. We were, however, able to discuss site activity with a group of the site's founders and long time users. Two of the subjects were founders and owners of the site. An additional two have been members for 7 years and are now site administrators. As long time users with high levels of access to the site, they were able to provide us with unique insight into which documents were the most important. For period 1, they speculated that Napster might be a major topic within the website, both because music downloads account for a large portion of Internet activity, and also because Napster was a hotly debated subject among the Internet community during period 1. Next, they suspected that http://everything2.org itself would be a major topic of user activity, as http://everything2.org has a fast growing user base who need to learn how to use the site and many existing users need to learn about the new or more advanced features of the site. Finally, they strongly believed that there would be sex-related topics in the website, as is common in many online communities without central control and close censoring.

For period 2, they suspected that Napster would be a less salient topic. Napster was effectively shut down prior to period 2, which began in November 2001, and people had gradually lost interest in the legal debate. http://everything2.org itself, however, would stay a major topic, for the community kept attracting new users. In addition, they believed that sex-related topics would still be prominent in period 2.

5.1.2. Evaluation of a priori hypotheses for period 1

As we have explained, the top n principal clusters are derived from the top n singular values and corresponding singular vectors using an SVD of the navigation matrix. For the navigation matrix from each time period, we computed the 10 largest singular values and associated singular vectors. For each singular vector we identified up to three positions with the largest nonzero loadings as hubs or authorities.

Table 2 lists the 1st, 2nd and the 7th principal clusters from our analysis for period 1. We named each principal cluster based on the contents of its hubs and authorities. Recall that we speculated that Napster would be a major topic. Napster turns out to be the No. 1 principal cluster. That is, the most prominent user activity in period 1 is about Napster. This cluster has a single hub "Napster" with loading 1, because the Napster node was the only nonzero value in the first left singular vector. The Napster node has several links to popular documents containing information on Napster and mp3 download, and, as our analysis detects, is the most likely starting point for a user with

Table 2
A few principal clusters in period 1

Principal cluster	Top hubs	*H*-scores	Top authorities	*A*-scores
1st, Napster	Napster	1	Napster of Puppets	0.999949
			Uberleech	0.007967
			Napigator	0.006137
2nd, How to use http://everything.org	Everything University	0.987253	Read me first	0.538687
	Tip of the day	0.150195	Everything is not a TV set	0.424028
	Read me first	0.030835	The newbie's guide...	0.267871
7th, Crime tutorials	Anarchist's cookbook	0.999986	Making plastic explosives from...	0.230248
	Ripping off soda machines	0.00297	Do ya hate school	0.216067
	How to annoy a fast-food...	0.002167	Ripping off change machines	0.2132

those interests. The top authority that we have identified for the same principal cluster, a node named "Napster of Puppets", is a popular destination for users who explore the Napster topic. Its loading is very close to 1, meaning that it is a very important destination on the Napster topic.

We also suspected that http://everything2.org itself would be a major topic of user activity. The 2nd, 3rd and 8th principal clusters turn out to be about how to use the http://everything2.org website, the philosophy of http://everything2.org, and how to gain "god" (power user) status in the community, respectively. Again confirming our suspicions, the 5th, 9th, 10th principal clusters turn out to be sex related. Due to the provocative titles of these documents, they are not described in this paper.

Note that the 7th cluster from period 1 is related to crimes. Some top authorities in that cluster, "How to make plastic explosives," and "Do ya hate school," are related to terrorism and hate crimes. As we mentioned before, because the community is open to everyone, http://everything2.org can possibly contain crime-related documents. While it is infeasible to manually censor all contributed content or monitor individual users' activities, our analysis seems to be able to uncover undesirable use of the system.

5.1.3. Evaluation of a priori hypotheses for period 2 and trend analysis

We also computed the top 10 principal clusters for the user navigations in period 2. A few of these principal clusters are shown in Table 3. As we suspected, Napster is no longer a major topic that appears among the top 10 clusters. Also as we suspected, there is one cluster (the 8th cluster) that is sex related. The http://everything2.org itself remains a major topic. In fact, the 1st, 2nd, 4th, 7th and 10th clusters are all about http://everything2.org. It is a little surprising to us that so many principal clusters are about the community itself. The 4th cluster, Content Rescue, and the 7th cluster, Node Tracker, are new topics that did not appear in top principal clusters of period 1. We closely examined these two principal clusters.

With more users writing new documents and appending to existing documents, the quality of many documents seems to have deteriorated. Similarly, the quality of some rosters (contact lists in instant

Table 3
A few principal clusters in period 2

Principal cluster	Top hubs	*H*-scores	Top authorities	*A*-scores
1st, How to use http://everything.org	Everything University	0.987113	The Newbie's Guide...	0.747613
	E2 HTML tags	0.131662	Words of advice for...	0.423053
			As Cool As It Gets	0.317936
4th, Content rescue	The Content Rescue	0.769443	Content Rescue: Nodes	0.924664
	Quests-Content Rescue	0.599463	Content Rescue: Roster	0.280511
	Content Rescue: Nodes	0.15408	Content Rescue: Darkroom	0.185671
7th, Node tracker	Node tracker	0.961507	E2 node tracker	0.997013
	Ak!You lost experience	0.174914	E2 is unfriendly to old noders	0.033878
	E2 Link/Logger Client	0.002167		

messaging) and discussion rooms suffered from increasing numbers of users. Content Rescue is a self-organized effort by http://everything2.org users to revive the quality of the documents, rosters, discussion rooms and other objects on the website.

http://everything2.org has a social status system that makes the community similar to a society. Through mechanism such as voting or earning points, users can gain social status by writing quality documents and actively participating in the community. The social status hierarchy is critical to the community's self-management. As we have seen in period 1, how to gain power or user status in the community was a prominent topic. Node tracker is an open source development effort to develop an automated agent that allows users to track their social status development in the community. For example, the agent can report to a user when other users have voted on the items he or she wrote. It appears that many users have contributed to development of this agent, and even more users are using this agent to track their social status development.

From the above, it appears that http://everything2.org has grown significantly as a community from period 1 to period 2. More user activities are community related, while overall the topics of user activities have become more diverse. Note that the "crime tutorials" topic in period 1 no longer appears in top principal clusters. Period 2 started 2 months after the September 11th tragedy. It appears that many users have shied away from terrorism or crime related topics.

5.1.4. External evaluation using votes, bookmarks and user interviews

Our results confirmed all of our a priori hypotheses about external events and behavior. To further evaluate our analysis, we used three additional external evaluation instruments: user votes on documents, user bookmarks and key informants' evaluation. In the following, we describe them in detail. For evaluation using user votes and bookmarks, we only report period 1 results because the evaluation for period 2 is similar.

In http://everything2.org, a logged-in user can vote on documents. Each vote is either a positive vote or a negative vote. A positive vote on a document, or the fact that a document is voted at all, indicates the importance of the document. For period 1, we collected 3,886,166 votes with 3,145,971 positives and 740,135 negatives. These votes are from 2,366 voters on 237,696 different documents. We code positive votes by 1 and negative votes by −1. Then for each voted document, we have a count of votes and a net total of votes. We assign unvoted documents both a count and a net total of zero.

Including unvoted documents, the mean vote count is 11.5 and the median is 3, with standard deviation of 51.5. As a contrast, for the hubs and authorities in top 10 principal clusters, the mean vote count is 116 and the median vote count is 55, with standard deviation of 22.4. The histogram in Fig. 3 compares the percentage of documents that have vote count in different ranges for all nodes and for top principal clusters. Principal clusters get more votes than average documents, with significance at a 10^{-4} level using t-test. In terms of median vote count, principal clusters rank in the top 4% among all the documents.

http://everything2.org also allows a user to maintain a list of shortcuts to frequently accessed documents. When a user reads a useful document, he or she can click on the "bookmark!" link in the right frame, which is the control frame available throughout the user's navigation within the site. A link to the bookmarked document is then added to the "personal bookmarks" section in the control frame. A user can create an unlimited number of bookmarks, although too many bookmarks may overcrowd the user's browser. In essence, bookmarking a document is one way to say that the document is very useful to the user.

For period 1, we collected 62,278 bookmarks created by 3864 users on 35,999 different documents.

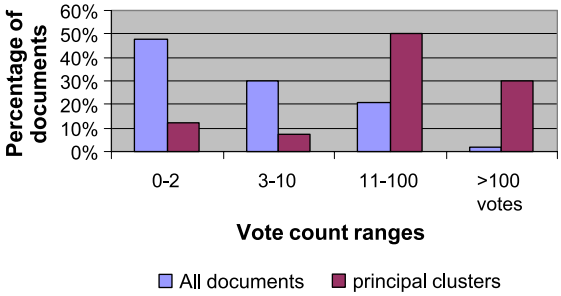


Fig. 3. Top 10 principal clusters get more votes than average documents do.

Only 10% of the documents have been bookmarked by any users, with 1.7 bookmarks on average (standard deviation: 2.86). In contrast, 45% of documents in top 10 principal clusters have been bookmarked, with 12 bookmarks on average. Including nonbookmarked documents, http://everything2.org gets 0.2 bookmarks per document, while top principal clusters get 4.7 bookmarks per document. Documents in principal clusters are 23 times more frequently bookmarked than average documents. Hubs appear to be more frequently bookmarked than authorities, although using *t*-test on the 10 top principal clusters we cannot confirm a difference at a 0.05 significance level. The histogram in Fig. 4 compares the percentage of documents that have bookmark counts in different ranges for all nodes and for top principal clusters.

It is clear that votes and bookmarks are not normally distributed across documents. We hypothesized that they also follow a power-law distribution. Using log-log linear regressions, we obtained the estimated vote distribution function as Probability (#vote+1=x)=0.512...$x^{-1.753}$. Similarly, the bookmark distribution function is estimated as Probability(#bookmarks+1=y)=0.128...$y^{-2.589}$. Note that it was necessary for us to add 1 to the number of votes and bookmarks, to accommodate zero values. Both estimates have R^2>0.9 which indicates a very good fit. The coefficients are all significant at a 10^{-6} level. We also used the curve estimation feature in SPSS™ to verify that the power function was the best fit among all available curve estimation functions in SPSS™. To our knowledge, our study is the first to find that users' votes and bookmarks on web documents follow power-law distributions.

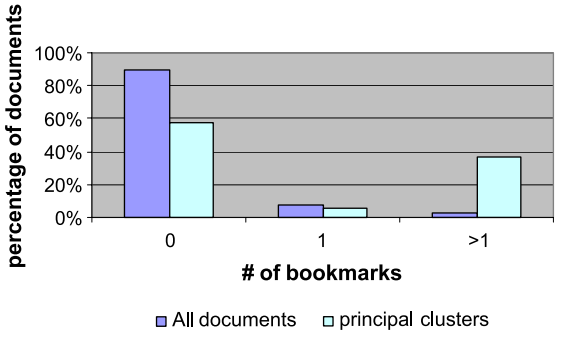


Fig. 4. Top 10 principal clusters get more bookmarks than average documents do.

There are some slight "disappointments" in our results. Why were only half of the hubs and authorities in principal clusters bookmarked? Why weren't they all highly voted? Why are many of the most requested documents not included in top principal clusters? By examining the content of the unbookmarked hubs and authorities, we found all of them were either sex or crime related. Users may have considered these documents inappropriate to be included in their personal bookmarks section, yet these documents are among the most popular, frequently accessed documents based on page request statistics. We also looked at the hubs and authorities that have less than 12 votes (where the mean vote count for all documents in the website was 11.5). We found these documents are all "factoid" style documents, which may be why users have not voted on them. Of the top 20 requested documents that are not included in the top 10 principal clusters, we found that they all belong to the "Napster" or "Everything2.org" topics. On the other hand, the top requested documents alone could only identify five out of 10 topics identified by our top 10 principal clusters. Overall then, our "disappointments" seem unfounded, and our findings seem to justify the use of Principal Clusters Analysis in addition to studying page request statistics, bookmarks and votes.

The group of site administrators interviewed for the initial hypotheses was also asked to evaluate the results of this analysis. We presented our principal clusters to them to see what they thought. Their feedback was very positive. We received comments such as " Those were all very popular nodes in the system at the time," and, "Some of the nodes are more practical than others, but they are all very highly rated in the system." The system administrators see the results as a useful, new way for them to discover major topics and patterns of user activity. The existing web server statistics tools is fairly rudimentary, with features such as a list of most frequently requested documents and recently created documents. But there are no tools to identify topics or the association among these documents. They are interested in collaborating with us to develop new website features such as document categorization and automated document rating. These key informants also reaffirmed some of our findings. They confirmed that people tend not to bookmark "inappropriate" documents despite frequent access.

Interviews also confirmed that nobody had an overall picture of user activities in this large community, including the system administrators. It is beyond human capacity to follow all activities in such a large environment with hundreds of thousands of users and documents. Most interviewees identified several principal clusters as prominent topics in the website, but none of them had intimate knowledge of even all topics in top 10 principal clusters. To any particular interviewee, many of these topics are "hidden" topics, although he or she spends a large amount of time daily on the website.

5.2. Internal criteria

To analyze further our methods, we sought to determine the degree to which we were able to capture both the structure of the entire document collection and the structure of each cluster.

For period 1, the matrix containing just the top 3 hubs and authorities based on each of the 10 largest singular values is a 16×24 matrix. What is contained in this matrix are the prominent activities within http://everything2.org and the key documents involved during the period. As shown in Fig. 5, the top 10 singular values account for 37.5% of total sum of squares of all singular values. Similar to the dimension reduction using principal component analysis, we have accounted for over 1/3 of the original navigation data by considering only 10 key underlying dimensions.

For period 2, the 10 largest singular values account for only 15% of total sum of squares. Note that the percentage of user activities accounted for by top 10 principal clusters is much smaller than in period 1. This indicates that user activity has become much more diverse, which seems to make sense as the number of documents and the number of users on the website both grew significantly during our data collection.

Key informants confirmed that, from period 1 to period 2, a large number of users besides the initial core group became deeply involved in the community development, including self-organized censoring and development of software tools. They also confirmed that the topics in the community had diversified significantly from period 1 to 2.

We were also interested in the coherence of the clusters we identified. The singular vectors from our calculation are orthonormalized. The sum of squares of loadings for the top n documents in a left or right singular vector indicates the degree to which the user activity in the corresponding principal cluster is centered on these documents. The closer the sum of squares of loadings is to one, the more centered the user activity. Fig. 6 illustrates the sums of squares of loadings for the top 3 hubs and authorities in each of the top 10 principal clusters. Note that the sums of squares of loadings for top 3 hubs are all close to 1, which indicates that there are usually a very small number of helpful starting points to explore a given topic. The sums of squares of loadings for top 3 authorities are sometimes much lower, which indicates that for some topics there may exist many useful destinations. Different

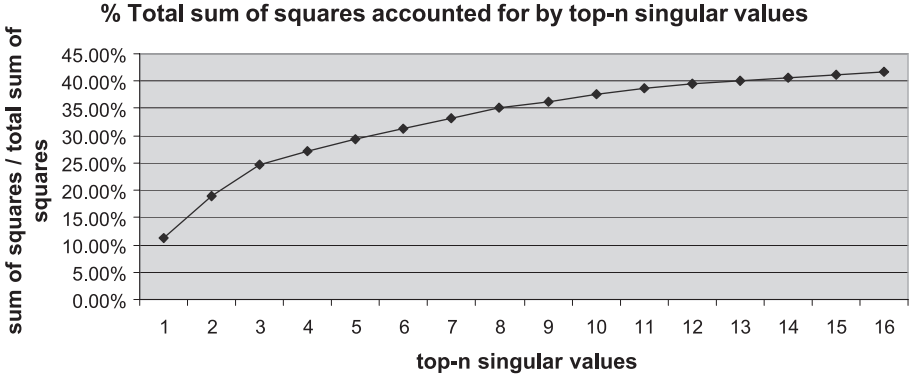


Fig. 5. The sums of squares of top n singular values divided by the total sum of squares indicate how closely the rank-n truncated SVDs approximate the original navigation matrix for period 1. Our choice of studying top 10 principal clusters seems to be reasonable, since the curve starts to flat off after first a few singular values.

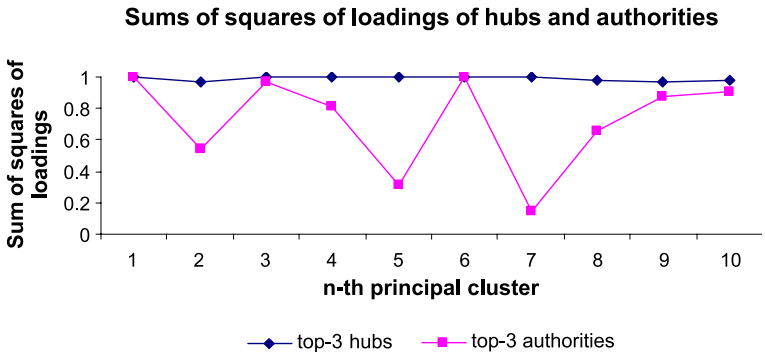


Fig. 6. The sum of squares of loadings of top 3 hubs and authorities for 10 principal clusters.

topics appear to affect the diversity of user activity. For Napster (cluster #1), a clearly defined topic, users' navigation is heavily clustered around one hub and three authorities. In other words, one hub and three authorities closely represent the entirety of user activity on this topic. Learning about http://everything2.org (2nd principal cluster) is a less clearly defined topic and the user navigation demonstrates less centrality. For this cluster, the squares of loadings of three hubs still sum close to 1, but the squares of loadings of three authorities sum only to 54%.

5.3. Replicabity

To test the robustness of our analysis, we created multiple random samples each containing 5% of the 13.5 million total navigations in period 1. We applied our principal clusters analysis to these samples and did not observe a significant difference in the resulting principal clusters (Table 4).

6. Discussion and conclusion

Our goal is to provide a summary of user navigation activities for intelligence workers' immediate consumption. We developed *principal clusters analysis*, a technique that successfully identifies a small number of major topics and a few important documents on these topics from a large amount of navigation data. The results are human-readable without the assistance of additional visualization or query tools. This is important since although computers can help process data, intelligence work ultimately requires human interpretation. We verified our mining results using various criteria and the results are consistent and very promising. To our knowledge, our study is the first to examine navigations in a large online community to obtain intelligence.

As an SVD-based technique, principal clusters analysis is scalable and independent of data collection techniques [4]. It holds the promise for many security or intelligence informatics applications. For example,

Table 4
A few principal clusters in period 1 using a 5% sample

Principal cluster	Top hubs	*H*-scores	Top authorities	*A*-scores
1st, Napster	Napster	1	Napster of Puppets	0.999705
			Uberleech	0.00917
			Napigator	0.00917
2nd, How to use http://everything.org	Everything University	0.986927	Read me first	0.5358
	Tip of the day	0.150302	Everything is not a TV set	0.425953
	Read me first	0.030901	The newbie's guide...	0.268485
7th, Crime tutorials	Anarchist's cookbook	0.999843	Making plastic explosives from...	0.222975
	Ripping off soda machines	0.003121	Do ya hate school	0.215724
	How to annoy a fast-food...	0.002186	Ripping off change machines	0.20851

The hubs/authorities in top 10 clusters are the same as the full data except that the loading scores are slightly different.

if authorized agencies monitor Internet traffic using router-based or snooper-based methods, our analysis technique would be able to process the data and distill major user activities. Our data analysis technique can be applied to data captured at a macro level such as people's navigations from website to website or from domain to domain. The results from our analysis allow analysts to take pinpointed actions. Recall that the 7th principal cluster of period 1 is on "crime tutorials". The top hub and authority in that clusters are "Anarchist's cookbook" and "Making plastic explosives", respectively. An effective way to disrupt propagation of such terrorism-related content is to remove the hubs and authorities related to this topic. Removing these articulation points is known to be the most effective way of breaking scale-free networks including terrorist networks [2].

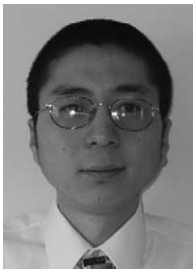
Our navigation mining is based on passive observation of user activities, which is more objective and comprehensive, yet less intrusive, than many other methods. We have found that many major user activities and important documents related to these activities are not revealed by users' explicit feedback such as votes, or their implicit feedback such as bookmarks. At the same time, we are successful in identifying popularly bookmarks whose storage on users' browsers would commonly hide them from view. We have also identified many topics "hidden" from any particular informant. Navigation mining seems to be a good way to complement intelligence collection through informants or group polls. As an alternative to examining usage patterns, one could conceive of analyzing the static hyperlink graph structures among all documents. While feasible (but still daunting) for a self-contained repository, this approach would only account for the website's static structure rather than users' behavior. Note that the documents may be dynamic, such as from dynamic queries to a database. A static snapshot of the website can only capture limited information and hyperlinks. In addition, navigations are not always performed through hyperlinks. Hence, the hyperlink graph would have a different structure than the navigation data we have collected.

We did not capture navigations of individuals or for a group of individuals due to privacy concerns. However, law enforcement and national security agencies would have the power to do so when necessary. Our data analysis technique can be applied to individuals or groups, if the identity of individuals can be identified in the web log. We believe that applying principal clusters analysis to individuals' or groups' navigation data would unveil more insightful information that may help promote security against terrorism.

References

[1] Access Log Analyzers. http://www.uu.se/Software/Analyzers/Access-analyzers.html, 2003.
[2] A.L. Barabasi, Linked: The New Science of Networks, Perseus, Cambridge, MA, 2002.
[3] M.W. Berry, Large scale sparse singular value decompositions, International Journal of Supercomputer Applications 6 (1) (1992) 33–49.
[4] M.W. Berry, S.T. Dumais, G.W. O'Brien, Using Linear Algebra for intelligent information retrieval, SIAM Review 37 (4) (1995) 177–196.
[5] L. Catledge, J. Pitkow, Characterizing Browsing Strategies in the World-Wide Web, Proc. 3rd International World Wide Web Conference, Darmstadt, Germany. 1995.
[6] M. Chen, A. LaPaugh, J.P. Singh, Predicting category accesses for a user in a structured information space, Proc. of SIGIR, 2002.
[7] C.W. Choo, B. Detlor, D. Turnbull, A behavioral model of information seeking on the web—preliminary results of a study of how managers and IT specialists Use the web, 1998 ASIS Annual Meeting, 1998.
[8] A. Cockburn, S. Jones, Which way now? Analysing and easing inadequacies in WWW navigation, International Journal of Human Computer Studies 45 (1) (1996) 129–205.
[9] R. Cooley, B. Mobasher, J. Srivastava, Web mining: information and pattern discovery on the World Wide Web, Proc. of the 9th IEEE International Conference on Tools with Artificial Intelligence, Dec, 1997.
[10] D. Crow, B. Smith, DB–Habits: comparing minimal knowledge and knowledge-based approaches to pattern recognition in the domain of user–computer interactions, in: Beale, Finlay (Eds.), Neural Networks and Pattern Recognition in Human–Computer Interaction, Ellis Horwood, New York, NY, 1992, pp. 39–61.
[11] J. Cugini, J. Scholtz, VISVIP: 3D Visualization of Paths through Web Sites, Proc. of International Workshop on Web-Based Information Visualization (WebVis'99), Florence, Italy, September 1–3, 1999, pp. 259–263.
[12] B. Davison, Web traffic logs: An imperfect resource for evaluation, Ninth Annual Conference of the Internet Society (INET'99). San Jose, 1999.
[13] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, R.A. Harshman, Indexing by Latent Semantic Analysis, JASIS 41 (6) (1990) 391–407.
[14] DVS-NetMon. http://www.dvsinfo.com., 2003.

[15] Ergosoft Laboratories, ergoBrowser. http://www.ergolabs.com/resources.htm, 2003.
[16] M. Etgen, J. Cantor, What does getting WET (Web Event-Logging Tool) mean for web usability? Fifth Human Factors and the Web Conference. 1999.
[17] A. Frieze, R. Kannan, S. Vempala, Fast Monte-Carlo algorithms for finding lowrank approximations, Proceedings of 37th Annual Symposium on Foundations of Computer Science, 1998.
[18] Y. Fu, K. Sandhu, M. Shih, Fast clustering of web users based on navigation patterns, Proc. of SCI/ISAS'99, 1999.
[19] J. Garofalakis, P. Kappos, D. Mourloukos, Web site optimization using page popularity, in: IEEE Internet Computing, vol. 3, 4, IEEE Computer Society, 1999 (July/August), pp. 22–29.
[20] M. Granovetter, The strength of weak ties, American Journal of Sociology 78 (6) (1973) 1360–1380.
[21] J.I. Hong, J. Heer, S. Waterson, J.A. Landay, WebQuilt: a proxy-based approach to remote web usability testing, ACM Transactions on Information Systems 19 (3) (2002) 263–285.
[22] X. Huang, N. Cercone, A. An, Comparison of interestingness functions for learning web usage patterns, ACM CIKM'02, Nov 4–9, 2002, Mclean, VA, 2002.
[23] T. Joachims, Optimizing search engines using clickthrough data, Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), ACM, 2002.
[24] J.D. Jobson, Applied Multivariate Data Analysis, Springer-Verlag, New York, 1992.
[25] K.P. Joshi, A. Joshi, Y. Yesha, Warehousing and mining web logs, Proc. of WIDM 1999, Kansas City, USA, 1999.
[26] J. Kleinberg, Authoritative sources in a hyperlinked environment, Proc. ACM-SIAM Symposium on Discrete Algorithms, 1998.
[27] S.R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, Trawling emerging cyber-communities automatically, Proc. 8th International World Wide Web Conferences, 1999.
[28] T. Landauer, S. Dumais, Solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge, Psychological Review 104 (2) (1997) 211–240.
[29] B. Mobasher, H. Dai, T. Luo, M. Nakagawa, Y. Sun, J. Wiltshire, Discovery of Aggregate Usage Profiles for Web Personalization, WEBKDD (2000).
[30] C. Papadimitriou, P. Raghavan, H. Tamaki, S. Vempala, Latent semantic indexing: a probabilistic analysis, Proc. of ACM Symposium on Principles of Database Systems, 1997.
[31] P. Pirolli, E. Pitkow, Distributions of surfers' paths through the World Wide Web: empirical characterization, World Wide Web 2 (1–2) (1999) 29–45.
[32] P. Pirolli, J. Pitkow, R. Rao, Silk from a sow's ear: extracting usable structures from the web, Proc. of the SIGCHI Conference on Human Factors in Computing Systems: Common Ground, April 13–18, 1996, Vancouver, Canada, ACM Press, 1996, pp. 118–125.
[33] M. Perkowitz, O. Etzioni, Towards adaptive web sites, Artificial Intelligence 118 (2000) 245–275.
[34] B. Prasetyo, I. Pramudiono, K. Takahashi, M. Kitsuregawa, Naviz: website navigational behavior visualizer, Proc. of 6th Pacific-Asia Conference Knowledge Discovery and Data Mining, Taipei, Taiwan, May 6–8, 2002.
[35] C. Shahabi, A. Zarkesh, J. Adibi, V. Shah, Knowledge discovery from users web-page navigation, Proc. of the IEEE 7th International Workshop on Research Issues in Data Engineering, 1997, pp. 20–30.
[36] M. Spiliopoulou, L.C. Faulstich, WUM: A Web Utilization Miner, EDBT Workshop WebDB98, Valencia, Spain, Springer Verlag, Valencia, Spain, 1998, pp. 184–203. http://www.springeronline.com/sgw/cda/frontpage/0,11855,4-40109-22-1642946-0,00.html.
[37] R. Srikant, R. Agrawal, Mining sequential patterns: generalizations and performance improvements, 5th Int'l Conference on Extending Database Technology, Avignon, France, March, 1996, pp. 3–17.
[38] R. Srikant, Y. Yang, Mining Web Logs to Improve Website Organization, Proc. of WWW10, 2001.
[39] Z. Su, Q. Yang, H.J. Zhang, X.W. Xu, Y.H. Hu, Correlation-based document clustering using web logs, Proc. of HICSS-34, 2001.
[40] L. Tauscher, S. Greenberg, Revisitation patterns in World Wide Web navigation, ACM SIG CHI 97, 1997.
[41] Vividence, Vividence CustomerScope™ http://www.vividence.com, 2003.
[42] J. Wang, C. Zheng, T. Li, W. Liu, Ranking user's relevance to a topic through link analysis on web logs, Proc. of ACM WIDM'02, Mclean, VA, 2002.
[43] H. Wu, M. Gordon, K. Demaagd, W. Fan, Principal clusters analysis, Proc. of the 18th International Conference of Information Systems (ICIS). Barcelona, Spain, December, 2002, pp. 757–762.
[44] R. Zaiane, M. Xin, J. Han, Discovering web access patterns and trends by applying OLAP and data mining technology on web logs, Proc. Advances in Digital Libraries Conf. (ADL'98), Santa Barbara, CA, April, 1998, pp. 19–29.
[45] J. Zhu, J. Hong, J. Hughes, Using Markov models for web site link prediction, Proc. of ACM Hypertext'02, ACM Press, 2002, pp. 131–139.



Harris Wu is an assistant professor at the College of Business and Public Administration, Old Dominion University. His research interests include knowledge management, information retrieval and information economics. He has published in leading journals such as Journal of the American Society for Information Science and Technology (JASIST), edited books and conference proceedings such as ACM/SIGCHI, ACM/Hypertext, ICIS and WWW. He has participated in research funded by Advanced Research and Development Activity in Information Technology, a US Government entity which sponsors and promotes research of import to the intelligence community which includes CIA, DIA, NSA, NGA and NRO. He has consulted at over a dozen of Fortune 500 companies and cofounded several IT businesses.

Michael Gordon is a professor of business information technology and associate dean for information technology at the University of Michigan Business School. His research interests include information retrieval, especially adaptive methods and methods that support knowledge sharing among groups; information and communication technology in the service of social enterprise (promoting economic development, providing health care delivery, and improving educational opportunities for the poor); and using information technology along with social methods to support business education.

Kurt DeMaagd is a PhD student at the University of Michigan Business School. His research interests include complex adaptive systems, multiagent modeling, supply chain management, information retrieval, online communities and open source software. His work has appeared in several edited books as well as in conferences such as the AAAI Symposium on Agent Mediated Knowledge Management, Communities and Technologies, ACM/SIGCHI, ACM/Hypertext, and ICIS. He is also active in the open source community. He was a cofounder of http://Slashdot.org and is a director of The Perl Foundation.

Weiguo Fan is an assistant professor of information systems and computer science at the Virginia Polytechnic Institute and State University. He received his PhD in Information Systems from the University of Michigan Business School, Ann Arbor, in July 2002. His research interests include personalization, data mining, text/web mining, web computing, business intelligence, digital library, and knowledge sharing and individual learning in online communities. His research has appeared in many prestigious information technology journals such as Information Processing and Management (IP and M), IEEE Transactions on Knowledge and Data Engineering (TKDE), Information Systems (IS), Decision Support Systems (DSS), ACM Transactions on Internet Technology (TOIT), Journal of the American Society for Information Science and Technology (JASIST), Journal of Classification, International Journal of Electronic Business, and in leading information technology conferences such as ICIS, HICSS, AMCIS, WWW, CIKM, DS, ICOTA etc.