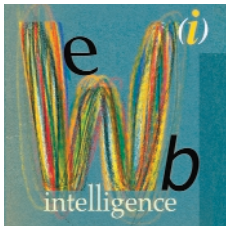
Data Mining for Web Intelligence



Data mining holds the key to uncovering and cataloging the authoritative links, traversal patterns, and semantic structures that will bring intelligence and direction to our Web interactions.

Jiawei Han
Kevin Chen-Chuan Chang
University of Illinois at Urbana-Champaign

Through the billions of Web pages created with HTML and XML, or generated dynamically by underlying Web database service engines, the Web captures almost all aspects of human endeavor and provides a fertile ground for data mining. However, searching, comprehending, and using the semistructured information stored on the Web poses a significant challenge because this data is more sophisticated and dynamic than the information that commercial database systems store.

To supplement keyword-based indexing, which forms the cornerstone for Web search engines, researchers have applied data mining to Web-page ranking. In this context, data mining helps Web search engines find high-quality Web pages[1] and enhances Web click stream analysis.[2]

For the Web to reach its full potential, however, we must improve its services, make it more comprehensible, and increase its usability. As researchers continue to develop data mining techniques, we believe this technology will play an increasingly important role in meeting the challenges of developing the intelligent Web.

WHY DATA MINING?

The Web—an immense and dynamic collection of pages that includes countless hyperlinks and huge volumes of access and usage information—provides a rich and unprecedented data mining source. However, the Web also poses several challenges to effective resource and knowledge discovery:

- *Web page complexity far exceeds the complexity of any traditional text document collection.* Although the Web functions as a huge digital library, the pages themselves lack a uniform structure and contain far more authoring style and content variations than any set of books or traditional text-based documents. Moreover, the tremendous number of documents in this digital library have not been indexed, which makes searching the data it contains extremely difficult.
- *The Web constitutes a highly dynamic information source.* Not only does the Web continue to grow rapidly, the information it holds also receives constant updates. News, stock market, service center, and corporate sites revise their Web pages regularly. Linkage information and access records also undergo frequent updates.
- *The Web serves a broad spectrum of user communities.* The Internet's rapidly expanding user community connects millions of workstations. These users have markedly different backgrounds, interests, and usage purposes. Many lack good knowledge of the information network's structure, are unaware of a particular search's heavy cost, frequently get lost within the Web's ocean of information, and can chafe at the many access hops and lengthy waits required to retrieve search results.
- *Only a small portion of the Web's pages contain truly relevant or useful information.* A given user generally focuses on only a tiny portion of the Web, dismissing the rest as uninteresting data that serves only to swamp the desired search results.

How can a search identify that portion of the Web that is truly relevant to one user's interests? How can a search find high-quality Web pages on a specified topic?

0018-9162/02/$17.00 © 2002 IEEE

Currently, users can choose from three major approaches when accessing information stored on the Web:

- *keyword-based search* or *topic-directory browsing* with search engines such as Google or Yahoo, which use keyword indices or manually built directories to find documents with specified keywords or topics;
- *querying deep Web sources*—where information, such as amazon.com's book data and realtor.com's real-estate data, hides behind searchable database query forms—that, unlike the surface Web, cannot be accessed through static URL links; and
- *random surfing* that follows Web linkage pointers.

The success of these techniques, especially with the more recent page ranking in Google and other search engines,[3] shows the Web's great promise to become the ultimate information system.

Design challenges

Defining how to design an intelligent Web presents a major research challenge. Achieving our vision of the Web's potential requires overcoming two fundamental problems.

First, at the abstraction level, the traditional schemes for accessing the immense amounts of data that reside on the Web fundamentally assume the text-oriented, keyword-based view of Web pages. We believe a data-oriented abstraction will enable a new range of functionalities. Second, at the service level, we must replace the current primitive access schemes with more sophisticated versions that can exploit the Web fully.

Access limitations

Although keyword-, address-, and topic-based Web search engines already support information searches, data mining will play an important role in Web intelligence because the Web's current incarnation still cannot provide high-quality, intelligent services. Several factors contribute to this problem and motivate our research.

Lack of high-quality keyword-based searches. The quality of keyword-based searches suffers from several inadequacies:

- a search often returns many answers, especially if the keywords posed include terms drawn from perennially popular categories such as sports, politics, or entertainment;
- overloading keyword semantics can return many low-quality answers—for example, depending on the context, a *jaguar* could be an animal, car, sports team, or computer; and
- a search can miss many highly related pages that do not explicitly contain the posed keywords—for example, a search for the term *data mining* can miss many highly regarded machine learning or statistical data analysis pages.

Incorporating data semantics could substantially enhance the quality of keyword-based searches.

Lack of effective deep-Web access. In July 2000, analysts estimated that searchable databases on the Web numbered at least 100,000. These databases provide high-quality, well-maintained information, but are not effectively accessible. Because current Web crawlers cannot query these databases, the data they contain remains invisible to traditional search engines.

In July 2000, analysts estimated that searchable databases on the Web numbered at least 100,000.

Conceptually, the deep Web provides an extremely large collection of autonomous and heterogeneous databases, each supporting specific query interfaces with different schema and query constraints. To effectively access the deep Web, we must integrate these databases.

Lack of automatically constructed directories. A topic- or type-oriented Web information directory presents an organized picture of a Web sector and supports semantics-based information searches—which makes such a directory highly desirable. For example, following hierarchical links like *US* > *universities* > *computer science* > *graduate program* makes searches more efficient. Unfortunately, developers must construct such directories manually. Even then, these costly directories provide only limited coverage and developers cannot easily scale or adapt them.

Lack of semantics-based query primitives. Most keyword-based search engines provide a small set of options for possible keyword combinations—essentially "with all the words" and "with any of the words." Some Web search services, such as Google and Yahoo, provide more advanced search primitives, including "with exact phrases," "without certain words," and with restrictions on date and domain site type.

Lack of feedback on human activities. Humanity's collective behavior often provides the best teacher. Web page authors provide links to "authoritative" Web pages and also traverse those Web pages they

Web page linkages contain many latent human annotations that can help automatically infer the notion of authority.

find most interesting or of highest quality.

Unfortunately, while human activities and interests change over time, Web links may not be updated to reflect these trends. For example, significant events—such as the 2002 World Cup finals or the terrorist attack of 11 September 2001—can change Web site access patterns dramatically, a change that Web linkages often fail to reflect. We have yet to use such human-traversal information for the dynamic, automatic adjustment of Web information services.

Lack of multidimensional analysis and data mining support. Because current Web searches rely on keyword-based indices, not the actual data the Web pages contain, search engines provide only limited support for multidimensional Web information analysis and data mining. For example, we cannot yet run queries that list major data mining research centers in North America, drill down through those sites that contain many research papers, then analyze the changes in their research focus based on these publications.

These challenges have promoted research into efficiently and effectively discovering and using Internet resources, a quest in which data mining will play an important role.

WEB MINING TASKS

The following tasks embody research problems that must be solved if we are to use data mining effectively in developing Web intelligence.

Mining Web search-engine data

An index-based Web search engine crawls the Web, indexes Web pages, and builds and stores huge keyword-based indices that help locate sets of Web pages that contain specific keywords. By using a set of tightly constrained keywords and phrases, an experienced user can quickly locate relevant documents.

However, current keyword-based search engines suffer from several deficiencies. First, a topic of any breadth can easily contain hundreds of thousands of documents. This can lead to a search engine returning a huge number of document entries, many of which are only marginally relevant to the topic or contain only poor-quality materials.

Second, many highly relevant documents may not contain keywords that explicitly define the topic, a phenomenon known as the *polysemy problem*. For example, the keyword *data mining* may turn up many Web pages related to other mining industries, yet fail to identify relevant papers on knowledge discovery, statistical analysis, or machine learning because they did not contain the data mining keyword.

Based on these observations, we believe data mining should be integrated with the Web search engine service to enhance the quality of Web searches. To do so, we can start by enlarging the set of search keywords to include a set of keyword synonyms. For example, a search for the keyword data mining can include a few synonyms so that an index-based Web search engine can perform a parallel search that will obtain a larger set of documents than the search for the keywords alone would return. The search engine then can search the set of relevant Web documents obtained so far to select a smaller set of highly relevant and authoritative documents to present to the user. Web-linkage and Web-dynamics analysis thus provide the basis for discovering high-quality documents.

Analyzing the Web's link structures

Given a keyword or topic, such as *investment,* we assume a user would like to find pages that are not only highly relevant, but authoritative and of high quality. Automatically identifying authoritative Web pages for a certain topic will enhance a Web search's quality.

The secret of authority hides in Web page linkages. These hyperlinks contain an enormous amount of latent human annotation that can help automatically infer the notion of authority. When a Web page's author creates a hyperlink pointing to another Web page, this action can be considered as an endorsement of that page. The collective endorsement of a given page by different authors on the Web can indicate the importance of the page and lead naturally to the discovery of authoritative Web pages. Thus the Web's linkage data provides a rich Web mining source. This idea has roots in traditional publishing as well: In the 1970s, researchers in information retrieval proposed methods for using journal article citations to evaluate the quality of research papers.[4] The Web linkage structure has several features that differ from journal citations, however.

First, not every hyperlink represents the endorsement a search is seeking. Web-page authors create some links for other purposes, such as navigation or to serve as paid advertisements. Overall, though, if most hyperlinks function as endorsements, the collective opinion will still dominate.

Second, an authority belonging to a commercial or competitive interest will seldom have its Web page point to rival authorities' pages. For example, Coca-Cola will likely avoid endorsing Pepsi by

ensuring that no links to Pepsi's Web pages appear on Coca-Cola's sites.

Third, authoritative pages seldom provide illuminating descriptions. For example, Yahoo's main Web page may not contain the explicit self-description "Web search engine."

These properties of Web link structures have led researchers to consider another important Web page category: *hubs*. A hub is a single Web page or page set that provides collections of links to authorities. Although it may not be prominent, or may have only a few links pointing to it, a hub provides links to a collection of prominent sites on a common topic.

These pages can be lists of recommended links on individual homepages, such as suggested reference sites from a course homepage or a professionally assembled resource list on a commercial site. A hub implicitly confers authority status on sites that focus on a specific topic. Generally, a good hub points to many good authorities, and, conversely, a page that many good hubs point to can be considered a good authority. Such a mutual reinforcement relationship between hubs and authorities helps users mine authoritative Web pages and automates discovery of high-quality Web structures and resources.

Methods for identifying authoritative Web pages and hubs have resulted in the development of the PageRank[1] and HITS[3] algorithms. Some commercially available Web search engines, such as Google, are built around such methods. By analyzing Web links and textual context information, these systems can generate better-quality search results than term-index engines such as AltaVista and topic directories such as Yahoo that human ontologists create.

Classifying Web documents automatically

Although Yahoo and similar Web directory service systems use human readers to classify Web documents, reduced cost and increased speed make automatic classification highly desirable. Typical classification methods use positive and negative examples as training sets, then assign each document a class label from a set of predefined topic categories based on preclassified document examples. For example, developers can use Yahoo's taxonomy and its associated documents as training and test sets to derive a Web document classification scheme. This scheme classifies new Web documents by assigning categories from the same taxonomy.[5]

Developers can obtain good results using typical keyword-based document classification methods—such as Bayesian classification, support vector machine, decision-tree induction, and keyword-based association analysis—to classify Web documents.[5,6] Since hyperlinks contain high-quality semantic clues to a page's topic, such semantic information can help achieve even better accuracy than that possible with pure keyword-based classification.

However, since the back-linked pages surrounding a document may be noisy and thus contain irrelevant topics, naive use of terms in a document's hyperlink neighborhood can degrade accuracy. For example, many personal homepages may have weather.com linked simply as a bookmark, even though these pages have no relevance to the topic of weather. Experiments have shown that coupling robust statistical models such as Markov random fields with relaxation labeling can substantially improve Web document classification accuracy.

Unlike many other classification schemes, automatic classification usually does not explicitly specify negative examples: We often only know which class a preclassified document belongs to, but not which documents a certain class definitely excludes. Thus, ideally, a Web document classification scheme should not require explicitly labeled negative examples. Using positive examples alone can be especially useful in Web document classification, prompting some researchers to propose a classification method based on a refined support-vector-machine scheme.[7]

Ideally, a Web document classification scheme should not require explicitly labeled negative examples.

Mining Web page semantic structures and page contents

Fully automatic extraction of Web page structures and semantic contents can be difficult given the current limitations on automated natural-language parsing. However, semiautomatic methods can recognize a large portion of such structures. Experts may still need to specify what kinds of structures and semantic contents a particular page type can have. Then a page-structure-extraction system can analyze the Web page to see whether and how a segment's content fits into one of the structures. Developers also can test user feedback to enhance the training and test processes and improve the quality of extracted Web page structures and contents.

Detailed analysis of Web page mining mechanisms reveals that different kinds of pages have different semantic structures. For example, a department's homepage, a professor's homepage, and a job advertisement page can all have different structures.

First, to identify the relevant and interesting structure to extract, either an expert manually specifies this structure for a given Web page class, or we develop techniques to automatically induce

Developers can use Web page structure and content extraction methods for automatic extraction based on Web page classes, possible semantic structures, and other semantic information.

such a structure from a set of prelabeled Web page examples. Second, developers can use Web page structure and content extraction methods[8] for automatic extraction based on Web page classes, possible semantic structures, and other semantic information. Page class recognition helps to extract semantic structures and contents, while extracting such structures helps to confirm which class the extracted pages belong to. Such an interaction mutually enhances both processes.

Third, semantic page structure and content recognition will greatly enhance the in-depth analysis of Web page contents and the building of a multilayered Web information base.

Mining Web dynamics

Web mining can also identify Web dynamics—how the Web changes in the context of its contents, structures, and access patterns.[2] Storing certain pieces of historical information related to these Web mining parameters aids in detecting changes in contents and linkages. In this case, we can compare images from different time stamps to identify the updates. However, unlike relational database systems, the Web's vast breadth and massive store of information make it nearly impossible to systematically store previous images or update logs. These constraints make detecting such changes generally infeasible. Mining Web access activities, on the other hand, is both feasible and, in many applications, quite useful.

With this technique, users can mine Web log records to discover Web page access patterns. Analyzing and exploring regularities in Web log records can enhance the quality and delivery of Internet information services to the end user, improve Web server system performance, and identify potential customers for electronic commerce.

A Web server usually registers a Web log entry for every Web page access. This entry includes the requested URL, the IP address from which the request originated, and a time stamp. Web-based e-commerce servers collect a huge number of Web access log records. Popular Web sites can register Web log records that number hundreds of megabytes each day. Web log databases provide rich information about Web dynamics. Accessing this information requires sophisticated Web log mining techniques.

The success of such applications depends on what and how much valid and reliable knowledge we can discover from the raw data. Often, researchers must clean, condense, and transform this data to retrieve and analyze significant and useful information. Second, researchers can use the available URL, time, IP address, and Web page content information to construct a multidimensional view on the Web log database and perform a multidimensional OLAP analysis to find the top users, top accessed Web pages, most frequently accessed time periods, and so on. These results will help discover potential customers, markets, and other entities.[9]

Third, mining Web log records can reveal association patterns, sequential patterns, and Web access trends. Web access pattern mining often requires taking further measures to obtain additional user traversal information. This data, which can include user browsing sequences from the Web server's buffer pages along with related data, facilitates detailed Web log analysis.

Researchers have used these Web log files to analyze system performance, improve system design through Web caching and page prefetching and swapping, determine the nature of Web traffic, and to evaluate user reaction to site design. For example, some studies have proposed adaptive Web sites that improve themselves by learning from user access patterns.[10]

Web log analysis can also help build customized Web services for individual users. Since Web log data provides information about specific pages' popularity and the methods used to access them, this information can be integrated with Web content and linkage structure mining to help rank Web pages, classify Web documents, and construct a multilayered Web information base.

Building a multilayered, multidimensional Web

We can construct and use the multidimensional Web in three major steps.

First, we systematically analyze a set of Web pages. This analysis, which covers page contents, structures, linkages, and usage patterns, can

- group a set of tightly integrated, closely related local Web pages into a cluster, called a semantic page; or
- treat an individual page, if it forms an independent cluster, as a semantic page.

The analysis then generates a descriptor for each semantic page, which contains a feature set critical for Web directory construction.

Second, we construct a semantics-based, evolving, multidimensional, multilayered Web information

directory, based on an expert-provided ontology and the semantic page descriptor database. We can use this directory system for query and information services, information analysis, and data mining.

Creating a Web warehouse that contains a copy of every page on the Internet would be unrealistic because it would require duplicating the entire Web. This insight indicates that the bottom layer of such a multilayered Web information structure cannot be a separate warehouse. This layer 0 must be the Web itself.

Layer 1, the Web page descriptor layer, contains descriptive information for all pages on the Web. Because layer 1 is an abstraction of layer 0, it should be substantially smaller, but it should still be rich enough to preserve most of layer 0's interesting, general information for keyword-based or multidimensional searches and mining. Based on the variety of Web page contents, layer 1 can be organized into dozens of semistructured classes, such as document, person, organization, advertisement, directory, sales, software, game, stocks, library catalog, and geographic and scientific data classes.

We can, for example, define the class *document* as follows:

> document (*file_addr, doc_category, authoritative_rank, key_words, authors, title, journal_or_book_title, publication_date, abstract, language, table_of_contents, categorydescription, index, links_out, multimediaattached, num_pages, form, size_doc, time_tamp, … , access_frequency*)

In this definition, each entry represents an abstraction of a Web page document. The first attribute, *file_addr*, registers the filename and the URL network address. The attributes *doc_category* and *authoritative_rank* contain crucial information that Web linkage analysis and document classification methods reveal. Many attributes—such as *key_words*, *authors*, *title*, *journal_or_book_title*, and so on—contain major semantic information that relates to the document. Other attributes provide formatting information, such as *form*, which indicates the file format. Several attributes register information directly associated with the file, such as *size_doc*, which lists the document file's size, and *time_stamp*, which notes when someone last modified the document.

We can construct various higher-layer Web directory services atop layer 1 to provide multidimensional, application-specific services. For example, we can construct yellow page services for database-system-oriented research. Such a directory could contain hierarchical structures for a few dimensions, such as theme category, geographical location, date of publication, and so on. Using Web page ranking and page or document classification services, we can choose to retain only information derived from relatively high-quality, highly relevant Web pages when we construct layer 1 or higher layers.

With the popular acceptance and adoption of XML, we anticipate that developers will use this structured markup language to write many future Web pages and, possibly, to share common Document Type Declarations.[11] This standardization would greatly facilitate information exchange among different Web sites and enhance information extraction for the construction of a multilayered Web information base. Further, we can design and implement Web-based information search and knowledge discovery languages specifically for this purpose.

Standardization would enhance information extraction for the construction of a multilayered Web information base.

It should thus be possible to construct a multilayered Web information base to facilitate resource discovery, multidimensional analysis, and data mining on the Internet—features that will likely form an important part of Internet-based information services.

Other data mining tasks for Web intelligence

Many other promising data mining methods can help achieve effective Web intelligence. Customizing service to a particular individual requires tracing that person's Web traversal history to build a profile, then providing intelligent, personalized Web services based on that information.

To date, some Web-based e-commerce service systems, such as amazon.com and expedia.com, register every user's past traversal or purchase history and build customer profiles from that data. Based on a user's profile and preferences, these sites select appropriate sales promotions and recommendations, thereby providing better quality service than sites that do not track and store this information. Using data mining to find a user's purchase or traversal patterns can further enhance these services.

Although a personalized Web service based on a user's traversal history could help recommend appropriate services, a system usually cannot collect enough information about a particular individual to warrant a quality recommendation. Either the traversal history has too little historical information about that person, or the possible spectrum of recommendations is too broad to set up a history

for any one individual. For example, many people make only a single book purchase, thus providing insufficient data to generate a reliable pattern.

In this case, collaborative filtering is effective because it does not rely on a particular individual's past experience but on the collective recommendations of the people who share patterns similar to the individual being examined. Thus, if people who have preferences similar to those of a given individual buy book A, they are likely to buy books B and C as well. The site could then recommend B and C to that individual. This approach generates quality recommendations by evaluating collective effort rather than basing recommendations on only one person's past experience. Indeed, collective filtering has been used as a data mining method for Web intelligence.[12]

Data mining for Web intelligence will be an important research thrust in Web technology—one that makes it possible to fully use the immense information available on the Web. However, we must overcome many research challenges before we can make the Web a richer, friendlier, and more intelligent resource that we can all share and explore. ■

References

1. S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Proc. 7th Int'l World Wide Web Conf.* (WWW98), ACM Press, New York, 1998, pp. 107-117.
2. J. Srivastava et al., "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," *SIGKDD Explorations*, vol. 1, no. 2, 2000, pp. 12-23.
3. S. Chakrabarti et al., "Mining the Web's Link Structure," *Computer,* Aug. 1999, pp. 60-67.
4. R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley, Reading, Mass., 1999.
5. S. Chakrabarti, *Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data*, Morgan Kaufmann, San Francisco, 2002.
6. J. Han and M. Kambert, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, 2001.
7. H. Yu, J. Han, and K.C-C. Chang, "PEBL: Positive Example-Based Learning for Web Page Classification Using SVM," *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery in Databases* (KDD02), ACM Press, New York, 2002, pp. 239-248.
8. V. R. Borkar, K. Deshmukh, and S. Sarawagi, "Automatic Segmentation of Text into Structured Records," *Proc. ACM-SIGMOD Int'l Conf. Management of Data (*SIGMOD 2001), ACM Press, New York, 2001, pp. 175-186.
9. S. Chaudhuri and U. Dayal, "An Overview of Data Warehousing and OLAP Technology," *SIGMOD Record*, vol. 26, no. 1, 1997, pp. 65-74.
10. M. Perkowitz and O. Etzioni, "Adaptive Web-Sites," *Comm. ACM,* vol. 43, no. 8, 2000, pp. 152-158.
11. S. Abiteboul, P. Buneman, and D. Suciu, *Data on the Web: From Relations to Semistructured Data and XML*, Morgan Kaufmann, San Francisco, 2000.
12. K. Yu et al., "Instance Selection Techniques for Memory-Based Collaborative Filtering," *Proc. SIAM Int'l Conf. Data Mining* (SIAM 02), ACM Press, New York, 2002, pp. 59-74.

Jiawei Han *is a professor in the Department of Computer Science, University of Illinois at Urbana-Champaign. His research interests include data mining, data warehousing, and knowledge discovery in databases. Han received a PhD in computer science from the University of Wisconsin at Madison. He is a member of the IEEE and the ACM. Contact him at hanj@cs.uiuc.edu.*

Kevin Chen-Chuan Chang *is an assistant professor in the Department of Computer Science, University of Illinois at Urbana-Champaign. His research interests include database systems, Internet information access, and information integration. Chang received a PhD in computer science from Stanford University. He is a member of the IEEE and the ACM. Contact him at kcchang@cs.uiuc.edu.*