CHAPTER-27
Mining the World Wide Web

27.Mining the World Wide Web

27.1 Introduction

The World Wide Web serves as a huge, widely distributed, global information service center for news, advertisements, consumer information, financial management, education, government, e-commerce, and many other information services. The Web also contains a rich and dynamic collection of hyperlink information and Web page access and usage information, providing rich sources for data mining. However, based on the following observations, the Web also poses great challenges for effective resource and knowledge discovery-

- The Web seems to be too huge/or effective data warehousing and data mining. The size of the Web is in the order of hundreds of terabytes and is still growing rapidly. Many organizations and societies put most of their public accessible information on the Web. It is barely possible to set up a data warehouse to replicate, store, or integrate all of the data on the Web.
- The complexity of Web pages is far greater than that of any traditional text document collection. Web pages lack a unifying structure. They contain far more authoring style and content variations than any set of books or other traditional text-based documents. The Web is considered a huge digital library, however, the tremendous number of documents in this library are not arranged according to any particular sorted order. There is no index by category, nor by title, author, cover page, table of contents, and so on. It can be very challenging to search for the information you desire in such a library!
- The Web is a highly dynamic information source: Not only does the Web grow at a rapid pace, its information is also constantly updated. News, stock markets, company advertisements, and Web service centers update their Web pages regularly. Linkage information and access records are also updated frequently. Recently, there have been efforts to store or integrate all of the data on the Web. For example, a huge Internet archive in the order of tens of terabytes can be accessed at http://www.archive.org/mdex.
- *The Web serves a broad diversity of user communities:* The Internet currently connects about 50 million workstations, and its user community is still expanding rapidly. Users may have very different backgrounds, interests, and usage purposes. Most users may not have good knowledge of the structure of the information network, may not be aware of the heavy cost of a particular search , may easily get lost by grouping in the “darkness” of the network, and may easily get bored by taking access “hops” and waiting impatiently for a piece of information.
- *Only a small portion of the information on the Web is truly relevant or useful:* It is said that 99% of the Web information is useless to 99% of Web users. Although this may not seem obvious, it is true that a particular person is generally interested in only a tiny portion of the Web, while the rest of the Web contains information that is uninteresting to the user and may swamp desired search results. How can the portion of the Web that is truly relevant to your interest be determined? How can we find high-quality Web pages on a specified topic?

These challenges have promoted research into efficient and effective discovery and use of resources on the Internet.

There are many index-based **Web search engines** that search the Web, index Web pages, and build and store huge keyword-based indices that help locate sets of Web pages containing certain keywords. With such search engines, an experiences user may be able to quickly locate documents by providing a set of tightly constrained keywords and phrases. However, current keyword-based search engines suffer from several deficiencies. First, a topic of any breadth may easily contain hundreds of thousands of documents. This can lead to a huge number of document entries returned by a search engine, many of which are only marginally relevant to the topic or may contain materials of poor quality. Second, many documents that are highly relevant to a topic may not contain keywords defining them. This is referred to as the polysemy problem, discussed in the previous section on text mining. 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 keyword data mining. As another example, a search based on the keyword search engine may not find even the most popular of the Web search engines like Yahoo!, Alta Vista, or America Online if these services do not claim to be search engines on their Web pages. This indicates that the current Web search engines are not sufficient for Web resource discovery.

"If Web search engines are not sufficient for Web resource discovery, how can we even think of doing Web mining?" Web mining is an even more challenging task that searches for Web access patterns, Web structures and the regularity and dynamics of Web contents. In general, Web mining tasks can be classified into three categories: Web contents mining, Web structure mining, and Web usage mining. Alternatively, Web structures can be treated as a part of Web contents so that Web mining can instead be simply classified into Web content mining and Web usage mining.

27.2 Mining the Web's Link Structure to Identify authoritative Web Pages

"What is meant by 'authoritative' Web pages?" Suppose you would like to search for Web pages relating to a given topic, such as financial investing. In addition to retrieving pages that are relevant, you also hope that the pages retrieved will be of high quality, or authoritative on the topic.

Interestingly, the secrecy of authority is hiding in Web page linkages. The Web consists not only of pages, but also of hyperlinks pointing from one page to another. These hyperlinks contain an enormous amount of latent human annotation that can help to automatically infer the notion of authority. When an author of a Web page creates a hyperlink pointing to another Web page, this can be considered as the author's endorsement of the other page. The collective endorsement of a given page by different authors on the Web may indicate the importance of the page and may naturally lead to the discovery of authoritative Web pages. Therefore, the tremendous amount of Web linkage information provides rich information about the relevance, the quality, and the structure of the Web's contents, and thus is a rich source for Web mining.

This ides has motivated some interesting studies on mining authoritative pages on the Web. In the 1970s, researches in information retrieval proposed methods of using citations among journal articles to evaluate the quality of research papers. However, unlike journal citations, the Web linkage structure has some unique features. First, not every hyperlink represents the endorsement we seek. Some links are created for other purposes, such as navigation or for paid advertisements. Yet overall, if majority of hyperlinks are for endorsement, the collective opinion will still dominate. Second, for commercial or competitive interests, one authority will seldom have its Web page point to its rival authorities in the same field. For example, Coca-Cola may prefer not to endorse its competitor Pepsi by not linking to

Pepsi's Web pages. Third, authoritative pages are seldom particularly descriptive. For example, the main Web pages of Yahoo! may not contain the explicit self-description "Web search engine".

These properties of Web link structures have led researches to consider another important category of Web pages called a hub. A hub is one or a set of Web pages that provides collections of rinks to authorities. Hub pages may not be prominent themselves, or there may exist few links pointing to them; however, they provide links to a collection of prominent sites on a common topic. Such pages could be lists of recommended links on individual home pages, such as recommended reference sites from a course home page, or professionally assembled resource lists on commercial sites. Hub pages play the role of implicitly conferring authorities on a focused topic. In general, a good hub is a page that points to many good authorities; a good authority is a page pointed to by many good hubs. Such a mutual reinforcement relationship between hubs and authorities helps the mining of authoritative Web pages and automated discovery of high quality Web structures and resources.

"So, how can we use hub pages to find authoritative pages?" An algorithm using hubs, called HITS (Hyperlink-Induced Topic Search), was developed as follows:

First, HITS uses the query terms to collect a starting set of, say, 200 pages from an index-based search engine. These pages form the root set. Since many of these pages are presumably relevant to the search topic, some of them should contain links to most of the prominent authorities. Therefore, the root set can be expanded into a base set by including all of the pages that the root-set pages link to, and all of the pages that link to a page m the root set, up to a designated size cutoff, such as 1000 to 5000 pages (to be included in the base set).

Second, a weight-propagation phase is initiated. This is an iterative process that determines numerical estimates of hub and authority weights. Notice that since the links between two pages with the same Web domain (i.e., sharing the same first level in their URLs) often serve as a navigation function and thus do not confer authority, such links are excluded from the weight-propagation analysis.

We first associate a nonnegative authority weight flp and a nonnegative hub weight hp with each page p in the base set, and initialize all a and h values to a uniform constant. The weights are normalized and an invariant is maintained that the squares of all weights sum to 1. The authority and hub weights are updated based on the following equations:

$$a_p = \sum h_q$$
$$h_p = \sum a_q$$

The first equation above implies that if a page is pointed to by many good hubs, its authority weight should increase (i.e., it is the sum of the current hub weights of all of the pages pointing to it). The second equation above implies that if a page is pointing to many good authorities, its hub weight should increase (i.e., it is the sum of the current authority weights of all of the pages it points to).

Finally, the HITS algorithm outputs a short list of the pages with large hub weights, and the pages with large authority weights for the given search topic. Many experiments have shown that HITS provides surprisingly good search results for a wide range of queries.

Although relying extensively on links can lead to encouraging results, the method may encounter some difficulties by ignoring textual contexts. For example, HITS sometimes drifts when hubs contain multiple topics. It may also cause "topic hijacking" when many pages from a single Web site point to the same single popular site, giving the site too large a share of the authority weight. Such problems can be overcome by replacing the sums of equations above with weighted sums, scaling down the weights of

multiple links from within the same site, using anchor text (the text surrounding hyperlink definitions in Web pages) to adjust the weight of the links along which authority is propagated, and breaking large hub pages into smaller units.

Systems based on the HITS algorithm include Clever and another system, Google, based on a similar principle. By analyzing Web links and textual context information, it has been reported that such systems can achieve better quality search results than those generated by term-index engines such as Alta Vista and those created by human ontologists such as Yahoo!.

27.3 Automatic Classification of Web Documents

In the automatic classification of Web documents, each document is assigned a class label from a set of predefined topic categories, based on a set of examples of pre-classified documents. For example, Yahoo's taxonomy and its associated documents can be used as training and test sets in order to derive a Web document classification scheme. This scheme may then be used to classify new Web documents by assigning categories from the same taxonomy.

Keyword-based document classification methods were the methods can be used for Web document classification. Such a term-based classification scheme has shown good results in Web page classification. Since hyperlinks contain high quality semantic clues to a page's topic, it is beneficial to make good use of such semantic information in order to achieve even better accuracy than pure keyword based classification. However, since the hyperlinks surrounding a document may be quite noisy, naïve use of terms in a document's hyperlink neighborhood can even degrade accuracy. The use of robust statistical models such as Markov random fields (MRFs), together with relaxation labeling, has been explored. Such a method has experimentally been shown to substantially improve the accuracy of Web document classification.

27.4 Construction of a Multilayered Web Information Base

A data warehouse can be constructed from a relational database to provide a multidimensional, hierarchical view of the data.

"Can we construct a multilayered Web information base to provide a multidimensional, hierarchical view of the Web?" you may wonder. Let us try to design such a multilayered Web information base to see whether it is realistic or beneficial.

First, it is unrealistic to create a Web warehouse containing a copy of every page on the Web, since this would just lead to a huge, duplicated WWW. This indicates that the bottom (most detailed) layer of such a multilayered Web information base must be the Web itself. It cannot be a separate warehouse. We will refer to this layer as layer-0.

Second, we can define layer-1 to be the Web page descriptor layer, containing descriptive information for pages on the Web. Hence, layer-1 is an abstraction of layer-0. It should be substantially smaller than layer-0 but still rich enough to preserve most of the interesting, general information for keyword-based or multidimensional search or mining.

Based on the variety of Web page contents, layer-1 can be organized into dozens of semi-structured classes, such as documents, person, organization, advertisement, directory, sales, software, game, stocks, library_catalog, geographic_data, scientific_data, and so on. For example, we may define class document as follows:

document(file_addr, doc_category, authoritative_rank, key_words, authors, title, publication, publication_date, abstract, language, table_of_contents, category_description, index, links_out, multimedia_attached, num_pages, form, size_doc, time_stamp,….,access_frequency),

where each entry is an abstraction of a document Web page. The first attribute, file_addr, registers the file name and the URL network address. The attributes doc_category and authoritative_rank contain crucial information that may be obtained by Web linkage analysis and document classification methods, as discussed in the previous two subsections. Many of the attributes contain major semantic information related to the document, such as key_words, authors, title, publication, publication_date, abstract, language, table_of_contents, index, links_out, multimedia_attached, and num_pages. Other attributes provide formatting information, such as form, which indicates the file format (e.g., ,ps, .pdf, .tex, .doc, .html, text, compressed, uuencoded, etc.). Several attributes register information directly associated with the file, such as size_doc (size of the document file) and time_stamp (time last modifies). The attribute access_frequency registers how frequently the entry is being accessed.

Third, various higher-layer Web directory services cab be constructed on top of layer-1 in order to provide multidimensional, application-specific services. For example, we may construct yellow page services for database-system-oriented research. Such a directory may contain hierarchical structures for a few dimensions, such as theme category, geographical location, date of publication, and so on.

"Do we really want to include information about every Web page?" Using Web page ranking and page or document classification services, we can choose to retain only the information necessary for relatively high-quality, highly relevant Web pages in the construction of layer-1 and/or higher layers of the information base.

With the popular acceptance and adoption of the structured Web page markup language, XML, it is expected that a large number of future Web pages will be written in XML and possibly share a good set of common DTDs (Document Type Declarations). Standardization with a language such as XML would greatly facilitate information exchange among different Web sites and information extraction for the construction of a multilayered Web information base. Furthermore, Web based information search and knowledge discovery languages can be designed and implemented for such a purpose.

In summary, based on the above, it should be possible to construct a multilayered Web information base to facilitate resource discovery, multidimensional analysis, and data mining on the Internet. It is expected that Web-based multidimensional analysis and data mining will form an important part of Internet-based information services.

27.5 Web Usage Mining

"What is Web usage mining?" Besides mining Web contents and Web linkage mines Web log records to discover user access patterns of Web pages. Analyzing and exploring regularities in Web log records can identify potential customers for electronic commerce, enhance the quality and delivery of Internet information services to the end user, and improve Web server system performance.

A Web server usually registers a (Web) log entry, or **Weblog entry**, for every access of a Web page. It includes the URL requested, the IP address from which the request originated, and a timestamp. For Web-based e-commerce servers, a huge number of Web access log records are being collected. Popular Web sites may register the Weblog records in the order of hundreds of megabytes every day. Weblog databases provide rich information about Web dynamics. Thus it is important to develop sophisticated Weblog mining techniques.

In developing techniques for Web usage mining, we may consider the following. First, although it is encouraging and exciting to imagine the various potential applications of Weblog file analysis, it is important to know that the success of such applications depends on what and how much valid and reliable knowledge can be discovered from the large raw log data. Often, raw Weblog data need to be cleaned, condensed, and transformed in order to retrieve and analyze significant and useful information. In principle, these preprocessing methods are similar to those discussed in chapter 3, although Weblog customized preprocessing is often needed.

Second, with the available URL, time, IP address, and Web page content information, a multidimensional view can be constructed on the Weblog database, and multidimensional OLAP analysis can be performed to find the top N users, top N accessed Web pages, most frequently accessed time periods, and so on, which will help discover potential customers, users, markets, and others.

Third, data mining can be performed on Weblog records to find association patterns, sequential patterns, and trends of Web accessing. For Web access pattern mining, it is often necessary to take further measures to obtain additional information of user traversal to facilitate detailed Weblog analysis. Such additional information may include user-browsing sequences of the Web pages in the Web server buffer, and so on.

With the use of such Weblog files, studies have been conducted on analyzing system performance, improving system design by Web caching. Web page pre-fetching, and Web pages swapping; understanding the nature of Web traffic and understanding user reaction and motivation. For example, some studies have proposed adaptive sites; Web sites that improve themselves by learning from user access patterns. Weblog analysis may also help build customized Web services for individual users.

Since Weblog data provide information about what kind of users will access what kind of Web pages, Weblog information can be integrated can be integrated with Web content and Web linkage structure mining to help Web page ranking, Web document classification, and the construction of a multilayered Web information base as well.

27.6 Review Questions

1 Explain about Mining the Web's Link Structure to Identify authoritative Web Pages

2 Explain about Automatic Classification of Web Documents

3Explain about Construction of a Multilayered Web Information Base

4 Explain about Web Usage Mining

27.7 References

[1]. Data Mining Techniques, Arun k pujari 1st Edition

[2] .Data warehousung,Data Mining and OLAP, Alex Berson ,smith.j. Stephen

[3].Data Mining Concepts and Techniques ,Jiawei Han and MichelineKamber

[4]Data Mining Introductory and Advanced topics, Margaret H Dunham PEA

[5] The Data Warehouse lifecycle toolkit , Ralph Kimball Wiley student Edition