14 Web Mining

Learning Objectives

* Understand the concept of Web Mining
* Know the goals and process of web content mining
* Know the goals and process of web structure mining
* Learn about hub and authority pages
* Understand the goals, architecture, and process of web usage mining
* Compute PageRank manually for a small network

### INTRODUCTION

Web Mining is the art and science of discovering patterns and insights from the World Wide Web to improve it. The World Wide Web is at the heart of the digital revolution. More data is posted on the web every day than was there on the whole web just 20 years ago. Billions of users are using it every day for a variety of purposes. The web is used for electronic commerce, business communication, and many other applications. Web mining analyzes data from the web and helps find insights that could optimize the web content and improve the user experience. Data for web mining is collected via web crawlers, weblogs, and other means.

Here are some characteristics of optimized websites

*Appearance* Aesthetic design,well-formatted content, easy to scan and navigate, good color contrasts

*Content* Well-planned information architecture with useful content, Fresh content, search engine optimized, links to other good sites

*Functionality* Accessible to all authorized users, fast loading times, usable forms, mobile-enabled

This type of content and its structure is of interest to ensure the web is easy to use. The analysis of web usage provides feedback on web content and also the consumer’s browsing habits. This data can be of immense use for commercial advertising and even for social engineering.

The web could be analyzed for its structure as well as the content. The usage pattern of web pages could also be analyzed. Depending upon objectives, web mining can

be divided into three different types – Web usage mining, Web content mining, and Web structure mining (Figure 14.1).



Web Mining

Web Content—

HTML Content

Web Structure—

URL Links

Web Usage—

Site visits, clicks

FIGURE 14.1 Web Mining Structure

### WEB CONTENT MINING

A website is designed in the form of pages with a distinct URL (Universal Resource Locator). A large website may contain thousands of pages. These pages (and their content) are managed using specialized software systems called Content Management Systems. Every page can have text, graphics, audio, video, forms, applications, and more kinds of content including user-generated content.

The websites keep a record of all requests received for their page/URLs, including the requester information using ‘cookies’. The log of these requests could be analyzed to gauge the popularity of those pages among different segments of the population. The text and application content on the pages could be analyzed for its usage by visit counts. The pages on a website themselves could be analyzed for the quality of content that attracts most users. Thus, the unwanted or unpopular pages could be weeded out or they can be transformed with different content and style. Similarly, more resources could be assigned to keep the more popular pages fresh and inviting.

### WEB STRUCTURE MINING

The web works through a system of hyperlinks using the hypertext protocol (HTTP). Any page can create a hyperlink to any other page, i.e., it can be linked to by another page. The intertwined or self-referral nature of the web lends itself to some unique network analytical algorithms. The structure of web pages could also be analyzed to examine the pattern of hyperlinks among pages. There are two basic strategic models for successful websites –Hubs and Authorities.

*Hubs* These are pages with a large number of interesting links. They serve as a hub or a gathering point, where people visit to access a variety of information.

Media sites like yahoo.com or government sites could serve that purpose. More focused sites like traveladvisor.com and yelp.com could aspire to become hubs for new emerging areas.

*Authorities* Ultimately, people would gravitate towards pages that provide the most complete and authoritative information on a particular subject. This could be factual information, news, advice, user reviews, etc. These websites have the most number of inbound links from other websites. Thus, mayoclinic.com could serve as an authoritative page for expert medical opinion; NYtimes.com could serve as an authoritative page for daily news.

### WEB USAGE MINING

As a user clicks anywhere on a webpage or application, the action is recorded by many entities in many locations. The browser at the client machine will record the click and the web server providing the content would also make a record of the pages served and the user activity on those pages. The entities between the client and the server, such as the router, proxy server, or ad server too would record that click.

The goal of web usage mining is to extract useful information and patterns from data generated through web page visits and transactions. The activity data comes from data stored in server access logs, referrer logs, agent logs, and client-side cookies. The user characteristics and usage profiles are also gathered directly or indirectly through syndicated data. Further, metadata such as page attributes, content attributes, and usage data are also gathered.

The web content could be analyzed at multiple levels (Figure 14.2).



Website

Visitors, Users, Customers

Web Logs

Web Click- streams

Gather and

Prepare Data for Analysis

–Identify users

–Identify visits

–Identify page views

–Summarize activity

Mine Data

for Patterns

–Usage patterns

–Usage profiles

–Web page profiles

–Web optimization

FIGURE 14.2 Web Usage Mining Architecture

* The *server-side analysis* would show the relative popularity of the web pages accessed. Those websites could be hubs and authorities.
* The *client-side analysis* would focus on the usage pattern or the actual content consumed and created by users.

1. Usage pattern could be analyzed using ‘clickstream’ analysis, i.e., analyzing web activity for patterns of a sequence of clicks, and the location and duration of visits on websites. Clickstream analysis can be useful for web activity analysis, software testing, market research, and analyzing employee productivity.
2. Textual information accessed on the pages retrieved by users could be analyzed using text mining techniques. The text would be gathered and structured using the bag-of-words technique to build a term-document matrix. This matrix could then be mined using cluster analysis and association rules for patterns such as popular topics, user segmentation, and sentiment analysis.

Web usage mining has many business applications. It can help predict user behavior based on previously learned rules and users’ profiles and can help determine the lifetime value of clients. It can also help design cross-marketing strategies across products by observing association rules among the pages on the website. Web usage can help evaluate promotional campaigns and see if the users were attracted to the website and used the pages relevant to the campaign. Web usage mining could be used to present dynamic information to users based on their interests and profiles. This includes targeted online ads and coupons at user groups based on user access patterns.

### WEB MINING ALGORITHMS

Hyperlink-Induced Topic Search (HITS) is a link analysis algorithm that rates web pages as being hubs or authorities. Many other HITS-based algorithms have also been published. The most famous and powerful of these algorithms is the PageRank algorithm. Invented by Google co-founder Larry Page, this algorithm is used by Google to organize the results of its search function. This algorithm helps determine the relative importance of any particular web page by counting the number and quality of links to a page. The websites with more number of links, and/or more links from higher-quality websites, will be ranked higher. It works similarly to determining the status of a person in a society of people. Those with relations to more people and/or relations to people of higher status will be accorded higher status.

PageRank is the algorithm that helps determine the order of pages listed upon a Google search query. The original PageRank algorithm formulation has been updated in many ways and the latest algorithm is kept a secret so other websites

cannot take advantage of the algorithm and manipulate their website according to it. However, many standard elements remain unchanged. These elements lead to the principles for a good website. This process is also called Search Engine Optimization (SEO).

## Conclusion

The web has been growing resources with more content each day and more users visiting it for many purposes. A good website should be useful, easy to use, and flexible for evolution. From the insights gleaned using web mining, websites should be constantly optimized.

Web usage mining can help discover what content users like and consume and help prioritize that for improvement. Web structure can help improve traffic to those sites, by building authority for the sites.

## Questions

1. What are the three types of web mining?
2. What is a clickstream analysis?
3. What are the two major ways by which a website can become popular?
4. What are the privacy issues in web mining?
5. A user spends 60 minutes on the web, visiting 10 webpages in all. Given the clickstream data, what kind of analysis would you do?

## True/False

1. The Web and the Internet are two names for the same thing.
2. The goal of web content mining is to look for broken URLs.
3. PageRank algorithm was created by Yahoo to determine the importance of webpages about the overall structure of the web.
4. Hub pages are well-known websites that provide definitive information on a particular set of topics.
5. NYtimes.com is an example of a hub page.
6. A cookie is a piece of data stored on a web browser to keep track of a user’s usage history.
7. Clickstream analysis is a part of web structure analysis.
8. All data mining techniques can be applied to web mining data.