Web Mining
Part 1

Dr. Sanjay Ranka
Manas Somaiya

Computer and Information Science and Engineering
University of Florida

Introduction

- The Web can be viewed as the largest *database* available and presents a challenging task for effective design and access
- Obviously the term database is used in a very loose context here as there is no structure or schema to the Web

Introduction …

- Data mining applied to the Web has been quite beneficial
- *Web Mining* is mining of data related to the World Wide Web
- This may be data actually present in Web pages or data related to Web activity

Web Data

- Web data
  - Content
  - Intra and Inter page structure
  - Usage data tha
  - User profiles)

Targeting

- One of the most important applications of Web mining is *targeting*
- Targeting is any technique that is used to direct business marketing or advertising to the most beneficial subset of total population
- The objective is to maximize the results of advertising; i.e. send it to all (and only) the set of potential customers who will buy
- The cost of sending an advertisement to someone who will not purchase that product can be avoided

Targeting …

- Targeting attempts to send advertisements to people who have not been to a Web site to entice them to visit it. Thus a targeted ad is found on a different Web site
- Targeting can be used to display advertising at Web sites visited by persons that fit in to a business' target demographic area
- By examining the Web log data to see what source sites access a Web site, information about the visitors can be obtained. This in turn can be used to sell advertising space to those companies that would benefit the most

Web Content Mining

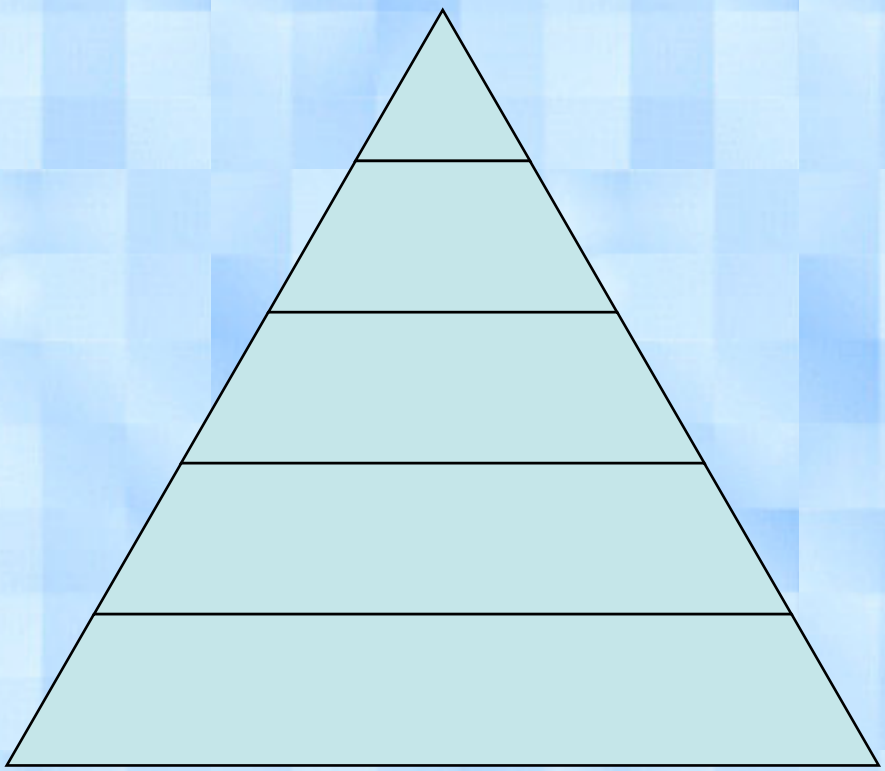
- Web content mining can be thought of as extending the work done by traditional search engines
- Most search engines are keyword based. Web content mining goes beyond this basic Information Retrieval technology
- It can improve on traditional search engines by using
  - Concept hierarchies and synonyms
  - User profiles
  - Analyzing links between pages

Web Content Mining …

- Traditional search engines have:
  - Crawlers: to search the Web and gather information
  - Indexes: various indexing techniques to store the information
  - Query Engines: query processing support in order to provide fast and accurate information to the end users
- Data mining techniques can be used to help search engines provide the efficiency, effectiveness and scalability needed

Web Content Mining …

- Basic content mining is a type of text mining
- A modified version of text mining functions [Zai99] can be viewed as a hierarchy with the simplest functions at the top and the more complex functions at the bottom



Web Content Mining …

- Many Web content mining activities are centered around techniques to summarize the information found
- Simple search engines use traditional IR technique. Their functionality could be extended to include more mining type activities

Crawlers

- A *robot* (or *spider* or *crawler*) is a program that traverses the hypertext structure in the Web
- The page (or a set of pages) that the crawler starts with are referred to as the *seed URLs*
- By starting at one page, all links from it are recorded and saved in a queue. These new pages are in turn searched and links are saved

Crawlers …

- As these robots search the Web, they might collect information about each page, such as extract key words and store in indices for the users of the associated search engine
- Crawlers are used to facilitate the creation of indices used by search engines. They allow the indices to be kept relatively up-to-date with little human intervention

Crawlers …

- A crawler may visit a certain number of pages and then stop, build an index, and replace the existing index. This type of crawler is referred to as *periodic crawler* because it is activated periodically
- Traditional crawlers usually replace the entire index or a section thereof. An incremental crawler selectively searches the Web and only updates the index incrementally as opposed to replacing it

Crawlers …

- A *focused crawler* visits pages related to topics of interest. With focused crawling, if it is determined that a page is not relevant or its links should not be followed, then the entire set of possible pages underneath it are pruned and not visited
- With thousands of focused crawlers, more of the Web can covered than with traditional crawlers. This facilitates better scalability as the Web grows

Personalization

- With *personalization,* Web access or the contents of a Web page are modified to better fit the desires of the user
- This may involve actually creating Web pages that are unique per user or using the desires of a user to determine what Web documents to retrieve
- The simplest example of personalization is the use of a visitor's name when he or she visits a page

Personalization …

- Personalization is almost the opposite of targeting. With targeting, businesses display advertisements at other sites visited by their potential customers. With personalization, when a particular person visits a Web site, the advertising can be designed specifically for that person
- The goal here is to entice a current customer to purchase something he or she may not have thought about purchasing
- For example, some Web sites allow personalization based on users' zip code

Personalization …

- Personalization includes techniques such as use of cookies, use of databases, and more complex data mining and machine learning strategies
- For example, a Web site may require that a visitor log on and provide information. This not only facilitates storage of personal information (by ID), but also avoids the problem of user identification with any type of Web mining
- Mining activities related to personalization may require examining Web log data to uncover patterns of access behavior by use. This may actually fall in to the category of Web usage mining

Personalization …

- Personalization can be viewed as a type of clustering, classification, or even prediction
- Through classification, the desires of a user are determined based on those for the class
- With clustering, the desires are determines based on those users to which he or she is determined to be similar
- Finally, prediction is used to predict what the user really wants to see

Personalization …

- There are three basic types of Web page personalization:
  - Manual techniques perform personalization through user registration preferences or via the use of rules that are used to classify individuals based on profiles or demographics
  - *Collaborative filtering* accomplishes personalization by recommending information (pages) that have previously been given high ratings by similar users
  - Content based filtering retrieves pages based on similarity between them and user profiles

Personalization …

- One of the earliest uses of personalization was with "My Yahoo!". Here a user himself personalizes what the screen looks like. He can provide preferences in areas such as weather, news, stock quotes, movies and sports
- Once the preferences are set up, each time the user logs in, his page is displayed. The personalization is accomplished by the user explicitly indicating what he wishes to see

Personalization …

- Some observations about the use of personalization with My Yahoo! are:
  - A few users will create very sophisticated pages by utilizing the customization provided
  - Most users do not seem to understand what personalization means and use only the default page
  - Any personalization system should be able to support both types of users
- This personalization is not automatic, but more sophisticated approaches to personalization actually use data mining techniques to determine the user preferences

Personalization …

- An automated personalization technique predicts future needs based on past needs or the needs of similar users
- Here interestingness is based on the similarity between the document and that of what the user wishes

Web Structure Mining

- Can be viewed as creating a model for the Web organization or a portion thereof
- This can be used to classify Web pages or to create similarity measures between documents

Page Rank

- The *Page Rank* technique was designed to both increase the effectiveness of search engines and improve their efficiency [PBMW98]
- Page Rank is used to measure the importance of a page and to prioritize pages returned from a traditional search engine using keyword searching
- The effectiveness of this measure has been demonstrated by the success of Google [Goo00]
- The name *Google* comes from the word *googol*, which is 10^{100}

Page Rank …

- The Page Rank value for a page is calculated based on the number of pages that point to it. This is actually measure based on the number of *backlinks* to a page
- The measure is not simply a count of the number of backlinks because a weighting is used to provide more importance to backlinks coming from important pages

Page Rank …

- Given a page p, we use B_p to be the set of pages that point to p, and F_p to be the set of links out of p. The Page Rank of a page p is defined as

$$PR(p) = c \sum_{q \in B_p} \frac{PR(q)}{N_q}$$

- Here $N_q = | F_q |$. The constant c is a value between 0 and 1 and is used for normalization

Clever

- Developed by IBM, *clever*, is aimed at finding both authoritative pages and hubs [CDK99]
- Defines an *authority* as the "best source" for the requested information. A *hub* is a page that contains links to authoritative pages
- Clever identifies authoritative pages and hub pages by creating weights
- A search can be viewed as having a goal of finding the best hubs and authorities

Clever …

- Because of the distributed and unsupervised development of sites, a user has no way of knowing whether the information contained within a Web page is accurate
- There is nothing to prevent someone from producing a page that contains not only errors, but also blatant lies
- Also some pages might be of higher quality than others. These pages are often referred to as being more authoritative
- Being authoritative is different than being relevant. A page may be extremely relevant but if it contains factual errors, users certainly do not want to retrieve it

Web Usage Mining

- Performs mining on Web usage data or Web logs
- A *Web log* is a listing of page reference data. Sometimes it is referred to as *clickstream* data because each entry corresponds to a mouse click
- These logs can be examined from either a client perspective or a server perspective
- When evaluated from a server perspective, mining uncovers information about the sites where the service resides. It can be used to improve the design of the sites
- By evaluating a client's sequence of clicks, information about a user (or group of users) is detected. This could be used to perform pre-fetching and caching of pages

HITS

- *Hyperlink-induced topic search (HITS)* finds hubs and authoritative pages [Kle99]
- The HITS technique contains two components:
  - Based on a given set of keywords (found in a query), a set of relevant pages is found
  - Hub and authority measures are associated with these pages. Pages with highest values are returned
- More details about the algorithm and an implementation of weight calculations using an adjacency matrix can be found in [Kle99]

Web Usage Mining …

- [SCDT00] identifies the following applications of Web usage mining:
  - Personalization for a user can be achieved by keeping track of previously accessed pages. These pages can be used to identify the typical browsing behavior of a user and subsequently to predict desired pages
  - By determining frequent access behavior of users, needed links can be identified to improve the overall performance of future accesses
  - Information about frequently accessed pages can be used for caching

Web Usage Mining …

- In addition to modifications to the linkage structure, identifying common access behaviors can be used to improve the actual design of the Web pages and to make other modifications to the site
- Web usage patterns can be used to gather business intelligence to improve sales and advertisement

Web Usage Mining …

- Web usage mining actually consists of three separate types of activities [SCDT00]:
  - Pre-processing activities center around reformatting the Web log data before it can mined
  - Pattern discovery activities from the major portion of the mining activities because these activities look to find hidden patterns within the log data
  - Pattern analysis is the process of looking at and interpreting the results of the discovery activities

Web Usage Mining …

- There are many issues associated with using the Web log for mining purposes:
  - Identification of the exact user is not possible from the log alone
  - With a Web client cache, the exact sequence of pages a user actually visits is difficult to uncover from the server site. Pages that are referenced may be found in the cache
  - There are many security, privacy and legal issues yet to be solved
    - Is the set of pages a person visits actually private information?
    - Should a Web browser actually divulge information to other companies about the habits of its users?