Web Mining
Part 1

Dr. Sanjay Ranka
Manas Somaiya

Computer and Information Science and Engineering
University of Florida

Introduction

- Determining the size of Web can be extremely difficult. In 2001, Google had announced that it indexes 3 billion Web documents
- 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 the potential to be 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 can be classified in to the following classes [SCDT00]:
  – Content of actual Web pages
  – Intra-page structure (includes the HTML or XML code for the page)
  – Usage data that describes how Web pages are accessed by visitors
  – User profiles (includes demographic and registration information obtained about users)
Taxonomy of Web Mining Activities

- Web mining tasks can be divided into several classes.
- One of the popular taxonomies is the one proposed by Zaiane [Zai99]

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.
- One problem associated with retrieval of data from Web documents is that they are not structured as in traditional databases. Once, HTML is replaced by XML, it will provide structured documents and facilitate easier mining.

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

• 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

• Recent research has examined how to use *incremental crawler*. 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 be covered than with traditional crawlers. This facilitates better scalability as the Web grows.

Crawlers …

• The focused crawler architecture consists of three primary components [CvdBD99]:
  – Hypertext classifier: associates a relevance score for each document with respect to a crawl topic; determines a resource rating that estimates how beneficial it would be to follow the links out of that page.
  – Distiller: determines which pages contain links to many relevant pages.
    • These are called hub pages. They are very important pages to be visited. They may not contain relevant information, but they would be quite important to facilitate continuing the search.
Crawlers …

– Crawler: performs the actual crawling on the Web. The pages it visits are determined via a priority-based structure governed by priority associated with the pages by the classifier and the distiller

• A performance objective for the focused crawler is a high precision or *harvest rate*

• More recent work on focused crawling has proposed the use of context graphs

Crawlers …

• Backward crawling finds pages that are not pointed to by relevant documents but are themselves relevant

• These types of pages may be new and may not yet have been discovered and linked to from other pages

• Although, backward links do not really exist in the Web, a backward crawl can be performed relatively easily because most search engines already maintain information about the back links

• This type of information is similar to that often used by commercial citation servers, which find documents that cite a given document. Some famous citation servers are
  – Science Citation Index
  – CiteSeer
Harvest System

- The Harvest system is based on the use of caching, indexing, and crawling
- Harvest is actually a set of tools that facilitate gathering of information from diverse sources
- The Harvest design is centered around the use of gatherers and brokers
- A gatherer obtains information for indexing from an Internet service provider, while a broker provides the index and query interface
- Brokers may interface directly with the gatherers or may go through other brokers to get to the gatherers

Harvest System …

- Harvest gatherers use the Essence system to assist in collecting data. Although not designed explicitly for use on the Web, Essence has been shown to be a valid technique for retrieving Web documents [HS93]
- Essence classifies documents by creating a semantic index.
- Semantic indexing generates different types of information for different types of files and then creates indices on this information.
- This process may first classify files based on type and then summarize the files typically based on keywords. Essence uses the file extensions to help classify file types
Virtual Web View

• One of the proposed approaches to handling the large amounts of unstructured data on the Web is to create a multiple layered database (MLDB) on top of the data in the Web (or a portion thereof).

• Such a database is massive and distributed in nature. Each layer of this database in more generalized than the layer beneath it.

• Unlike the lowest level (the Web), the upper layers are structured and can be mined by an SQL-like query language.

Virtual Web View ...

• The MLDB provides an abstracted and condensed view of a portion of the Web. A view of the MLDB, which is called a Virtual Web View (VWV), can be constructed.

• The indexing approach used by MLDB does not require the use of spiders.

• The Web servers themselves send their indices (or changes to indices) to the site(s) where indexing is being performed. This process is triggered when changes to the sites are made.
Virtual Web View …

• To assist in the creation of the first layer of the MLDB, both extraction and translation tools are proposed
  – Translation tools are used to convert Web documents to XML
  – Extraction tools extract the desired information from the Web pages and inserts it in to the first layer of the MLDB
  – It is proposed that any translation functions be performed directly by the local administration

Virtual Web View …

• The layer-1 data can be viewed as a massive distributed database. The higher levels of the database become less distributed and more summarized as they move up the hierarchy
• Generalization tools are proposed, and concept hierarchies are used to assist in the generalization process for constructing the higher levels of the MLDB. These hierarchies can be constructed using the WordNet Semantic Network
• WordNet is a database of the English language. Nouns, adjectives, verbs and adverbs are listed, divided in to groups of synonyms, and linked together using both lexical and semantic relationships
Virtual Web View …

• A Web data mining query language, WebML is proposed to provide data mining operations on the MLDB. Documents are accessed using data mining operations and lists of keywords.

• A major feature of WebML is four primitive operations based on the use of concept hierarchies for the keywords [Zai99]

Virtual Web View …

• These four primitives are:
  – COVERS: One concept covers another if it is higher in the hierarchy. This coverage is extended to include synonyms as well.
  – COVERED BY: This is the reverse of COVERS in that it reverses to descendents.
  – LIKE: The concept is a synonym.
  – CLOSE TO: One concept is close to another if it is a sibling in the hierarchy. Again this is extended to include synonyms.
Virtual Web View

- The following WebML query finds all documents at the level of “www.ufl.edu” that have a keyword that covers the keyword cat:

```
SELECT *
FROM document in “www.ufl.edu”
WHERE ONE OF keywords COVERS “cat”
```

- Data mining functions such as classification, summarization, association rules, clustering and prediction are included in WebML

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 determined 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
- *News Dude* uses the *interestingness* of a document to determine if a user is interested in it [BP99]
- Here interestingness is based on the similarity between the document and that of what the user wishes

Personalization …

- Similarity is measured by co-occurrence of words in the documents and a user profile created for the user
- The target application for News Dude is news stories
- News Dude actually prunes out stories that are too close to the stories that the user has already seen. These are determined to be redundant articles
Personalization …

• News Dude uses a two-level scheme to determine interestingness. One level is based on recent articles the user has read, while the second level is a more long-term profile of general interest.
• Thus, a short-term profile is created that summarizes recently read articles, and a long-term profile is created to summarize general interests. A document is found to be interesting if it is sufficiently close to either.
• It was shown that the use of this two-level approach works better than either profile by itself [MPR00].

Personalization …

• Another approach to automatic personalization is that used by *Firefly*.
• Firefly is based on the concept that humans often base decisions on what they hear from others. If someone likes a TV show, a friend of that person might also like that program.
• User profiles are created by users indicating their preferences.
Personalization …

- Prediction of a user’s desire is then made based on what similar users like. This approach to Web mining is called **collaborative filtering**
- The initial application of Firefly has been to predict music that a user would like
- Note that there is no examination of the actual Web content of documents, simply a prediction based on what similar users like

Personalization …

- Another collaborative approach is called **Web Watcher**.
- Web Watcher prioritizes links found on a page based on a user profile and the results of other users with similar profiles who have visited this page [JFM97]
- A user is required to indicate the intent of the browsing session. This profile is then matched to the links the user follows
References

- BP99, D. Billsus and M. Pazzani. A hybrid user model for news story classification
- CvdBD99, Soumen Chakrabarti, Martin van den Berg, and Byron Dorm. Focused Crawling: A new approach to topic-specific Web resource discovery
- HS93, Darren R. Hardy and Michael F. Schwartz. Essence: A resource discovery system based on semantic file indexing
- MPR00, Udi Manber, Ash pater, and John Robinson. Experience with personalization on yahoo!
- SCDT00, Jaideep Srivastava, Robert Cooley, Mukund Deshpande and Pang-Ning Tang. Web usage mining: Discovery and applications of usage patterns from Web data
- Zai99, Osmar Rachid Zaiane. Resource and knowledge discovery from the internet and multimedia repositories