Web Mining

What is Web Mining?

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents/services (Etzioni, 1996, CACM 39(11))

What is Web Mining?

- Motivation / Opportunity
  - The WWW is huge, widely distributed, global information service centre and, therefore, constitutes a rich source for data mining
    - Personalization, Recommendation Engines
    - Web-commerce applications
    - Building the Semantic Web
    - Intelligent Web Search
    - Hypertext classification and Categorization
    - Information / trend monitoring
    - Analysis of online communities

The Web

- Over 1 billion HTML pages, 15 terabytes
- Wealth of information
  - Bookstores, restaurants, travel, malls, dictionaries, news, stock quotes, yellow & white pages, maps, markets, ........
  - Diverse media types: text, images, audio, video
  - Heterogeneous formats: HTML, XML, postscript, pdf, JPEG, MPEG, MP3
- Highly Dynamic
  - 1 million new pages each day
  - Average page changes in a few weeks
- Graph structure with links between pages
  - Average page has 7-10 links
  - in-links and out-links follow power-law distribution
- Hundreds of millions of queries per day
Abundance and authority crisis

- Liberal and informal culture of content generation and dissemination
- Redundancy and non-standard form and content
- Millions of qualifying pages for most broad queries
  - Example: java or kayaking
- No authoritative information about the reliability of a site
- Little support for adapting to the background of specific users

How do you suggest we could estimate the size of the web?

One Interesting Approach

- The number of web servers was estimated by sampling and testing random IP address numbers and determining the fraction of such tests that successfully located a web server
- The estimate of the average number of pages per server was obtained by crawling a sample of the servers identified in the first experiment


The Web

- The Web is a huge collection of documents except for
  - Hyper-link information
  - Access and usage information
- Lots of data on user access patterns
  - Web logs contain sequence of URLs accessed by users
- **Challenge:** Develop new Web mining algorithms and adapt traditional data mining algorithms to
  - Exploit hyper-links and access patterns
Applications of web mining

- E-commerce (Infrastructure)
  - Generate user profiles -> improving customization and provide users with pages, advertisements of interest
  - Targeted advertising -> Ads are a major source of revenue for Web portals (e.g., Yahoo, Lycos) and E-commerce sites. Internet advertising is probably the "hottest" web mining application today
  - Fraud -> Maintain a signature for each user based on buying patterns on the Web (e.g., amount spent, categories of items bought). If buying pattern changes significantly, then signal fraud
- Network Management
  - Performance management -> Annual bandwidth demand is increasing ten-fold on average, annual bandwidth supply is rising only by a factor of three. Result is frequent congestion. During a major event (World cup), an overwhelming number of user requests can result in millions of redundant copies of data flowing back and forth across the world
  - Fault management -> analyze alarm and traffic data to carry out root cause analysis of faults

Why is Web Information Retrieval Important?

- According to most predictions, the majority of human information will be available on the Web in ten years
- Effective information retrieval can aid in
  - Research: Find all papers about web mining
  - Health/Medicine: What could be reason for symptoms of "yellow eyes", high fever and frequent vomiting
  - Travel: Find information on the tropical island of St. Lucia
  - Business: Find companies that manufacture digital signal processors
  - Entertainment: Find all movies starring Marilyn Monroe during the years 1960 and 1970
  - Arts: Find all short stories written by Jhumpa Lahiri

Why is Web Information Retrieval Difficult?

- The Abundance Problem (99% of information of no interest to 99% of people)
  - Hundreds of irrelevant documents returned in response to a search query
- Limited Coverage of the Web (Internet sources hidden behind search interfaces)
  - Largest crawlers cover less than 18% of Web pages
- The Web is extremely dynamic
  - Lots of pages added, removed and changed every day
- Very high dimensionality (thousands of dimensions)
- Limited query interface based on keyword-oriented search
- Limited customization to individual users

Information retrieval (Search) on the Web

- Automated generation of topic hierarchies
- Web knowledge bases
Search Engine Relative Size

Total Hits from 25 Searches
Dec. 31, 2002 ©2003 G. Nettles

- Google
- AlltheWeb
- AltaVista
- WiseNu
- HotBot
- MSN Search
- Toma
- NUREsearch
- Gigablast

<table>
<thead>
<tr>
<th>Engine</th>
<th>Hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>6133</td>
</tr>
<tr>
<td>AlltheWeb</td>
<td>4977</td>
</tr>
<tr>
<td>AltaVista</td>
<td>4664</td>
</tr>
<tr>
<td>WiseNu</td>
<td>3529</td>
</tr>
<tr>
<td>HotBot</td>
<td>3267</td>
</tr>
<tr>
<td>MSN Search</td>
<td>3267</td>
</tr>
<tr>
<td>Toma</td>
<td>3259</td>
</tr>
<tr>
<td>NUREsearch</td>
<td>2352</td>
</tr>
<tr>
<td>Gigablast</td>
<td>895</td>
</tr>
</tbody>
</table>

Coverage – about 40% in 1999

From http://www.searchengineshowdown.com/stats/size.shtml

Search Engine Web Coverage Overlap

Overlap of 4 Searches
141 Pages on Mar. 6, 2002

- Coverage – about 40% in 1999

End Of Size Wars? Google Says Most Comprehensive But Drops Home Page Count

- http://searchenginewatch.com/searchday/article.php/3551586
- By Danny Sullivan, Editor, September 27, 2005

- How do you measure Comprehensiveness?
  - Rare words
  - The Duplicate Content Issue
  - Counting Pages Indexed Per Site

Web Mining Taxonomy

Web Mining
- Web Content Mining
- Web Structure Mining
- Web Usage Mining
Web Mining Taxonomy

- **Web content mining**: focuses on techniques for assisting a user in finding documents that meet a certain criterion (text mining)
- **Web structure mining**: aims at developing techniques to take advantage of the collective judgement of web page quality which is available in the form of hyperlinks
- **Web usage mining**: focuses on techniques to study the user behaviour when navigating the web (also known as Web log mining and clickstream analysis)

Web Content Mining

Examines the content of web pages as well as results of web searching.

Web Content Mining

Can be thought of as extending the work performed by basic search engines

Search engines have **crawlers** to search the web and gather information, **indexing techniques** to store the information, and **query processing support** to provide information to the users

Web Content Mining is: the process of extracting knowledge from web contents

Semi-Structured Data

- Content is, in general, *semi-structured*
- Example:
  - Title
  - Author
  - Publication_Date
  - Length
  - Category
  - Abstract
  - Content

Structured attribute/value pairs

Unstructured
Structuring Textual Information

- Many methods designed to analyze structured data
- If we can represent documents by a set of attributes we will be able to use existing data mining methods
- How to represent a document?
  - Vector based representation
    - (referred to as "bag of words" as it is invariant to permutations)
  - Use statistics to add a numerical dimension to unstructured text

![Diagram of term frequency, document frequency, term proximity, and document length]

Document Representation

- A document representation aims to capture what the document is about
- One possible approach:
  - Each entry describes a document
  - Attribute describe whether or not a term appears in the document

Example

<table>
<thead>
<tr>
<th></th>
<th>Camera</th>
<th>Digital</th>
<th>Memory</th>
<th>Pixel</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document 1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Document 2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

- But a term is mentioned more times in longer documents
- Therefore, use relative frequency (% of document):
  - No. of occurrences/No. of words in document

Example: Term frequency table

<table>
<thead>
<tr>
<th>Terms</th>
<th>Camera</th>
<th>Digital</th>
<th>Memory</th>
<th>Print</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Terms</th>
<th>Camera</th>
<th>Digital</th>
<th>Memory</th>
<th>Print</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.03</td>
<td>0.02</td>
<td>0</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.004</td>
<td>0</td>
<td>0.003</td>
<td></td>
</tr>
</tbody>
</table>
More on Document Representation

- **Stop Word removal**: Many words are not informative and thus irrelevant for document representation.
  - the, and, a, an, is, of, that, …
- **Stemming**: reducing words to their root form (Reduce dimensionality).
  - A document may contain several occurrences of words like fish, fishes, fisher, and fishers.
  - But would not be retrieved by a query with the keyword fishing.
  - Different words share the same word stem and should be represented with its stem, instead of the actual word.
  - Fish.
- For the Portuguese language these techniques are less studied.

Weighting Scheme for Term Frequencies

- **TF-IDF weighting**: give higher weight to terms that are rare.
  - **TF**: term frequency (increases weight of frequent terms).
  - If a term is frequent in lots of documents it does not have discriminative power.
  - **IDF**: inverse term frequency.

For a given term \( w_j \) and document \( d_i \):

\[
\text{TF}_{ij} = \frac{n_{ij}}{|d_i|}
\]

\[
\text{IDF}_j = \log \frac{n}{n_j}
\]

\[
x_j = \text{TF}_{ij} \cdot \text{IDF}_j
\]

There is no compelling motivation for this method but it has been shown to be superior to other methods.

Locating Relevant Documents

- **Given a set of keywords**.
- **Use similarity/distance measure to find similar/relevant documents**.
- **Rank documents by their relevance/similarity**.

Distance Based Matching

- In order retrieve documents similar to a given document we need a measure of similarity.
- **Euclidean distance** (example of a metric distance): the distance between \( X=(x_1, x_2, x_3, \ldots, x_n) \) and \( Y=(y_1, y_2, y_3, \ldots, y_n) \) is defined as:

\[
D(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}
\]

**Properties of a metric distance**:
- \( D(X,Y)=D(Y,X) \)
- \( D(X,Z)+D(Z,Y) \geq D(X,Y) \)
**Angle Based Matching**

- Cosine of the angle between the vectors representing the document and the query.
- Documents "in the same direction" are closely related.
- Transforms the angular measure into a measure ranging from 1 for the highest similarity to 0 for the lowest.

\[
D(X,Y) = \cos(X,Y) = \frac{X^T Y}{\|X\| \cdot \|Y\|} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \cdot \sqrt{\sum y_i^2}}
\]

**Performance Measure**

- The set of retrieved documents can be formed by collecting the top-ranking documents according to a similarity measure.
- The quality of a collection can be compared by the two following measures:

  \[
  \text{precision} = \frac{|\text{Relevant} \cap \text{Retrieved}|}{|\text{Retrieved}|} \quad \text{percentage of retrieved documents that are in fact relevant to the query (i.e., "correct" responses)}
  \]

  \[
  \text{recall} = \frac{|\text{Relevant} \cap \text{Retrieved}|}{|\text{Relevant}|} \quad \text{percentage of documents that are relevant to the query and were, in fact, retrieved}
  \]

**Text Mining**

- Document classification
- Document clustering
- Key-word based association rules

**Web Search**

- Domain-specific search engines
  - www.buildingonline.com
  - www.lawcrawler.com
  - www.drkoop.com (medical)
- Meta-searching
  - Connects to multiple search engines and combine the search results
  - www.metacrawler.com
  - www.dogpile.com
  - www.37.com
Web Search

- Post-retrieval analysis and visualization
  - www.vivisimo.com
  - www.tumba.pt
  - www.kartoo.com
- Natural language processing
  - www.askjeeves.com
- Search Agents
  - Instead of storing a search index, search agents can perform real-time searches on the Web.
  - Fresher data, slower response time and lower coverage.

Focused Crawling

Breadth-first crawl

Focused crawl

Threshold: page is on-topic if correlation to the closest centroid is above this value.
Cutoff: follow links from pages whose "distance" from closest on-topic ancestor is less than this value

Database Approaches

- One approach is to build a local knowledge base - model data on the web and integrate them in a way that enables specifically designed query languages to query the data
- Store locally abstract characterizations of web pages. A query language enables to query the local repository at several levels of abstraction. As a result of the query the system may have to request pages from the web if more detail is needed

Agent-Based Approach

- Agents to search for relevant information using domain characteristics and user profiles
- A system for extracting a relation from the web, for example, a list of all the books referenced on the web. The system is given a set of training examples which are used to search the web for similar documents. Another application of this tool could be to build a relation with the name and address of restaurants referenced on the web


Web Structure Mining

Exploiting Hyperlink Structure

First generation of search engines

- Early days: keyword based searches
  - Keywords: "web mining"
  - Retrieves documents with "web" and mining"
- Later on: cope with
  - synonymy problem
  - polysemy problem
  - stop words
- Common characteristic: Only information on the pages is used

Modern search engines

- Link structure is very important
  - Adding a link: deliberate act
  - Harder to fool systems using in-links
  - Link is a "quality mark"

- Modern search engines use link structure as important source of information

Central Question:
Which useful information can be derived from the link structure of the web?
Some answers

1. Structure of Internet
2. Google
3. HITS: Hubs and Authorities

1. The Web Structure

- A study was conducted on a graph inferred from two large Altavista crawls.


- The study confirmed the hypothesis that the number of in-links and out-links to a page approximately follows a Zipf distribution (a particular case of a power-law)

Power Laws

\[ y = ax^k \]

\[ \log(y) = k \log(x) + \log(a) \]

In-Links
1. The Web Structure

- If the web is treated as an undirected graph 90% of the pages form a single connected component.
- If the web is treated as a directed graph four distinct components are identified, the four with similar size.

Some statistics

- Only between 25% of the pages there is a connecting path.
- But if there is a path:
  - Directed: average length <17
  - Undirected: average length <7 (!!!)
- It’s a “small world” -> between two people only chain of length 6!
- Small World Graphs
  - High number of relatively small cliques
  - Small diameter
  - Internet (SCC) is a small world graph

General Topology

- SCC: set of pages that can be reached by one another.
- IN: pages that have a path to SCC but not from it.
- OUT: pages that can be reached by SCC but not reach it.
- TENDRILS: pages that cannot reach and be reached the SCC pages.
2. Google

- Search engine that uses link structure to calculate a quality ranking (PageRank) for each page
- Intuition: PageRank can be seen as the probability that a “random surfer” visits a page
  

- Keywords \( w \) entered by user
- Select pages containing \( w \) and pages which have in-links with caption \( w \)
  
  - Anchor text
    - Provide more accurate descriptions of Web pages
    - Anchors exist for un-indexable documents (e.g., images)
  
  - Font sizes of words in text: Words in larger or bolder font are assigned higher weights

- Rank pages according to importance

PageRank

(PageRank) + (Website Content) = Overall Rank in Results

Page Rank: A page is important if many important pages link to it.

- Link \( i \rightarrow j \):
  - \( i \) considers \( j \) important.
  - The more important \( i \), the more important \( j \) becomes.
  - If \( i \) has many out-links: links are less important.

- Initially: all importances \( p_i = 1 \). Iteratively, \( p_i \) is refined.

\[
\text{PageRank}(j) = p \cdot (1-p) \sum \frac{\text{PageRank}(i)}{\text{OutDegree}(i)}
\]

PageRank

- Let OutDegree\(_i\) = \# out-links of page \( i \)
- Adjust \( p_j \):

\[
\text{PageRank}(j) = p \cdot (1-p) \sum \frac{\text{PageRank}(i)}{\text{OutDegree}(i)}
\]

This is the weighted sum of the importance of the pages referring to \( P_j \)

- Parameter \( p \) is probability that the surfer gets bored and starts on a new random page
- \((1-p)\) is the probability that the random surfer follows a link on current page

\[
\text{Parameter } p = 0.85
\]

Page D

Page C

Page B

Page A

Repeat until pagerank vector converges...
3. HITS (Hyperlink-Induced Topic Search)

- HITS uses hyperlink structure to identify authoritative Web sources for broad-topic information discovery.


- Premise: Sufficiently broad topics contain communities consisting of two types of hyperlinked pages:
  - Authorities: highly-referenced pages on a topic
  - Hubs: pages that "point" to authorities
  - A good authority is pointed to by many good hubs; a good hub points to many good authorities

Hubs and Authorities

- Hub pages point to interesting links to authorities = relevant pages
- Authorities are targets of hub pages

HITS

- Steps for Discovering Hubs and Authorities on a specific topic
  - Collect seed set of pages S (returned by search engine)
  - Expand seed set to contain pages that point to or are pointed to by pages in seed set (removes links inside a site)
  - Iteratively update hub weight $h(p)$ and authority weight $a(p)$ for each page:
    $$a(p) = \sum q h(q) \quad h(p) = \sum a(q)$$
  - After a fixed number of iterations, pages with highest hub/authority weights form core of community
  - Extensions proposed in Clever
    - Assign links different weights based on relevance of link anchor text

Applications of HITS

- Search engine querying
- Finding web communities
- Finding related pages
- Populating categories in web directories.
- Citation analysis
**Web Usage Mining**

analyzing user web navigation

---

**Web Usage Mining**

- Pages contain information
- Links are “roads”
- How do people navigate over the Internet?
- ⇒ Web usage mining (Clickstream Analysis)

- Information on navigation paths is available in log files.
- Logs can be examined from either a client or a server perspective.

---

**Website Usage Analysis**

- Why analyze Website usage?
- Knowledge about how visitors use Website could
  - Provide guidelines to web site reorganization; Help prevent disorientation
  - Help designers place important information where the visitors look for it
  - Pre-fetching and caching web pages
  - Provide adaptive Website (Personalization)
- Questions which could be answered
  - What are the differences in usage and access patterns among users?
  - What user behaviors change over time?
  - How usage patterns change with quality of service (slow/fast)?
  - What is the distribution of network traffic over time?
Data Sources

- **Server level collection**: the server stores data regarding requests performed by the client, thus data regard generally just one source;

- **Client level collection**: it is the client itself which sends to a repository information regarding the user's behaviour (can be implemented by using a remote agent (such as Javascripts or Java applets) or by modifying the source code of an existing browser (such as Mosaic or Mozilla) to enhance its data collection capabilities.);

- **Proxy level collection**: information is stored at the proxy side, thus Web data regards several Websites, but only users whose Web clients pass through the proxy.

An Example of a Web Server Log

Analog - Web Log File Analyser

Gives basic statistics such as
- number of hits
- average hits per time period
- what are the popular pages in your site
- who is visiting your site
- what keywords are users searching for to get to your site
- what is being downloaded
Web Usage Mining Process

Web Server Log → Data Preparation → Clean Data → Data Mining → Usage Patterns

Data Preparation

- Data cleaning
  - By checking the suffix of the URL name, for example, all log entries with filename suffixes such as, gif, jpeg, etc
- User identification
  - If a page is requested that is not directly linked to the previous pages, multiple users are assumed to exist on the same machine
  - Other heuristics involve using a combination of IP address, machine name, browser agent, and temporal information to identify users
- Transaction identification
  - All of the page references made by a user during a single visit to a site
  - Size of a transaction can range from a single page reference to all of the page references

Sessionizing

Main Questions:
- how to identify unique users
- how to identify/define a user transaction

Problems:
- user ids are often suppressed due to security concerns
- individual IP addresses are sometimes hidden behind proxy servers
- client-side & proxy caching makes server log data less reliable

Standard Solutions/Practices:
- user registration - practical ????
- client-side cookies - not fool proof
- cache busting - increases network traffic

Sessionizing

- Time oriented
  - By total duration of session
    - not more than 30 minutes
  - By page stay times (good for short sessions)
    - not more than 10 minutes per page
- Navigation oriented (good for short sessions and when timestamps unreliable)
  - Referrer is previous page in session, or
  - Referrer is undefined but request within 10 secs, or
  - Link from previous to current page in web site
- The task of identifying the sequence of requests from a user is not trivial - see Berendt et.al., Measuring the Accuracy of Sessionizers for Web Usage Analysis SIAM-DM01
**Web Usage Mining**

- Commonly used approaches
  - Preprocessing data and adapting existing data mining techniques
    - For example association rules: does not take into account the order of the page requests
  - Developing novel data mining models

**Association Rules**

- Find frequent patterns/associations/correlations among sets of items
- Find correlations between pages not directly connected
- Reveal associations between groups of users with specific interests
  - e.g.: `/events/ski.html, travel/ski_resorts.html →` /equipment/ski_boots.html (85%, 3%)
Mining Navigation Patterns

- Each session induces a user trail through the site.
- A trail is a sequence of web pages followed by a user during a session, ordered by time of access.
- A pattern in this context is a frequent trail.
- Co-occurrence of web pages is important, e.g. shopping-basket and checkout.
- Use a Markov chain model to model the user navigation records, inferred from log data.

Ngram model

- We make use of the Ngram concept in order to improve the model accuracy in representing user sessions.

The Ngram model assumes that only the previous n-1 visited pages have a direct effect on the probability of the next page chosen.
- A state corresponds to a navigation trail with n-1 pages.
- Chi-square test is used to assess the order of the model
  (in most cases N=3 is enough).
- Experiments have shown that the number of states is manageable.

First-Order Model

Input Streams

| A, B, C | A, B, D |
| A, B, C | E, B, D |
| E, B, C | E, B, D |

Number of traversals

Artificial state
First-Order Model

Input Streams

A,B,C  
A,B,D  
A,B,C  
E,B,D  
E,B,C  
E,B,D

Input Streams

A,B,C  
A,B,D  
A,B,C  
E,B,D  
E,B,C  
E,B,D

First-Order Model

Input Streams

A,B,C  
A,B,D  
A,B,C  
E,B,D  
E,B,C  
E,B,D

Input Streams

A,B,C  
A,B,D  
A,B,C  
E,B,D  
E,B,C  
E,B,D

Second Order Evaluation

Input Streams

A,B,C  
A,B,D  
A,B,C  
E,B,D  
E,B,C  
E,B,D

Transition probability

P(C | AB) = 0.67 Not accurate
Cloning

Duplicate states to separate in-links whose second-order probability diverge.

(state B cloned based on link A,B)

Clustering-Based Cloning

- In cases when a state has more than two in-links we use clustering for assigning in-links to clones.

- We use a state accuracy parameter, which sets the maximum admissible difference between corresponding first and second-order probabilities.

Clustering-Based Cloning

Input Streams

<table>
<thead>
<tr>
<th>Session</th>
<th>Occur.</th>
<th>2nd O. Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A,B,C</td>
<td>6</td>
<td>P(C</td>
</tr>
<tr>
<td>A,B,D</td>
<td>3</td>
<td>P(D</td>
</tr>
<tr>
<td>E,B,C</td>
<td>7</td>
<td>P(E</td>
</tr>
<tr>
<td>E,B,D</td>
<td>4</td>
<td>P(E</td>
</tr>
<tr>
<td>G,B,C</td>
<td>4</td>
<td>P(G</td>
</tr>
<tr>
<td>G,B,D</td>
<td>7</td>
<td>P(G</td>
</tr>
<tr>
<td>H,B,D</td>
<td>6</td>
<td>P(D</td>
</tr>
</tbody>
</table>

Not accurate
Clustering-Based Cloning

Applications of Markov models

- Provide guidelines for the optimisation of a web site structure.
- Work as a model of the user's preferences in the creation of adaptive web sites.
- Improve search engine's technologies by enhancing the random surf concept.
- Web personal assistant.
- Visualisation tool
- Use model to learn access patterns and predict future accesses. Pre-fetch predicted pages to reduce latency.
- Also cache results of popular search engine queries.

Summary

- Web is huge and dynamic
- Web mining makes use of data mining techniques to automatically discover and extract information from Web documents/services
  - Web content mining
  - Web structure mining
  - Web usage mining
- Semantic web: "The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." -- Tim Berners-Lee, James Hendler, Ora Lassila

References

- Data Mining: Introductory and Advanced Topics, Margaret Dunham (Prentice Hall, 2002)
- Mining the Web - Discovering Knowledge from Hypertext Data, Soumen Chakrabarti, Morgan-Kaufmann Publishers
Thank you !!!