Web Mining

Based on several presentations found on the web: Shapiro, Ullman, Terziyan, Pedersen, Bing Liu ...

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))

- Web mining aims to discovery useful information or knowledge from the Web hyperlink structure, page content and usage data.
  - (Bing LIU 2007, Web Data Mining, Springer)

Motivation / Opportunity

- The WWW is huge, widely distributed, global information service centre and, therefore, constitutes a rich source for data mining
  - Intelligent Web Search
  - Personalization, Recommendation Engines
  - Web-commerce applications
  - Building the Semantic Web
  - Web page classification and categorization
  - News classification and clustering
  - Information / trend monitoring
  - Analysis of online communities
  - Web and mail spam filtering

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
- Pages added continuously and average page changes in a few weeks
Table 1. A classification of retrieval and mining techniques and applications.

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Any data</th>
<th>Textual data</th>
<th>Web-related data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieving known data or documents efficiently and effectively</td>
<td>Data Retrieval / Database</td>
<td>Information Retrieval</td>
<td>Web Retrieval</td>
</tr>
<tr>
<td>Finding new patterns or knowledge previously unknown to the system</td>
<td>Data Mining</td>
<td>Text Mining</td>
<td>Web Mining</td>
</tr>
</tbody>
</table>

- **Number of pages**
  - Technically, infinite
    - Because of dynamically generated content
    - Lots of duplication (30-40%)
  - Best estimate of “unique” static HTML pages comes from search engine claims
    - Yahoo = claimed 19.2 billion in Aug 2005
- **Number of unique web sites**
  - Netcraft survey says 98 million sites

Another way to estimate the web size

- 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

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

Web Crawling Basics

Start with a “seed set” of to-visit urls

- get next url
- get page
- extract urls
- to visit urls
- visited urls
- web pages

Crawling Issues

- Load on web servers
  - E.g., no more than 1 request to the same server every 10 seconds
- Insufficient resources to crawl entire web
  - Visit “important” pages first (pagerank, inlinks ...)
- How to keep crawled pages “fresh”?
  - How often do web pages change? What do we mean by freshness?
- Detecting replicated content e.g., mirrors
  - Use document comparison techniques (java manuals)
- Can’t crawl the web from one machine
  - Parallelizing the crawl
Web Mining Taxonomy

- **Web content mining**: focuses on techniques for assisting a user in finding documents that meet a certain criterion
- **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.

- 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
Information Retrieval

Semi-Structured Data
- Text 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.

Document Representation
- One possible approach (boolean representation):
  - Each entry describes a document.
  - Attribute describe whether or not a term appears in the document.
  - Adequate for boolean queries.

| Terms | Camera | Digital | Memory | Pixel | ...
|-------|--------|---------|--------|-------|-------
| Document 1 | 1      | 1       | 0      | 1     | ...
| Document 2 | 1      | 1       | 0      | 0     | ...
| ...    | ...    | ...     | ...    | ...   | ...   |
Boolean Queries

- Each document or query is treated as a “bag” of words or terms. Word sequence is not considered.
- Query terms are combined logically using the Boolean operators AND, OR, and NOT.

Retrieval
- Given a Boolean query, the system retrieves every document that makes the query logically true.
- Called exact match.
- The retrieval results are usually quite poor because term frequency is not considered.

More on Document Representation

- Stop Word removal: Many words are not informative and thus irrelevant for document representation.
- Many of the most frequently used words in English are useless in IR and text mining – these words are called stop words.
  - the, and, a, an, is, of, that, ...
  - Typically about 400 to 500 such words
  - For an application, an additional domain specific stopwords list may be constructed
  - stopwords are not useful for searching or text mining
  - they may also confuse the retrieval system.

More on Document Representation

- 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

- Improving effectiveness of IR and text mining
  - matching similar words
  - Mainly improve recall

- For the Portuguese language these techniques are less studied

Document Representation

- Another approach:
  - Each entry describes a document
  - Attributes represent the frequency in which a term appears in the document

- Example: Term frequency table

| Terms  | Camera | Digital | Memory | Print | ...
|--------|--------|---------|--------|-------|-----
| Document 1 | 3      | 2       | 0      | 1     |     
| Document 2 | 0      | 4       | 0      | 3     |     
| ...     | ...    | ...     | ...    | ...   | ... |
Document Representation

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

| Terms       | Camera | Digital | Memory | Print | ...
|------------|--------|---------|--------|-------|------
| Document 1 | 0.03   | 0.02    | 0      | 0.01  | ...
| Document 2 | 0      | 0.004   | 0      | 0.003 | ...
| ...        | ...    | ...     | ...    | ...   | ...

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 $w_i$ and document $d_j$:
- $n_j$ is the number of occurrences of $w_i$ in document $d_j$
- $|d_j|$ is the number of words in document $d_j$
- $n$ is the number of documents
- $n_j$ is the number of documents that contain $w_i$

$$TF_i = \frac{n_i}{|d_j|}$$

$$IDF_j = \log \frac{n}{n_j}$$

$$x_i = TF_i \cdot IDF_j$$

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

How to determine if two documents are similar?

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 Euclidean distance between $X=(x_1, x_2, x_3,...,x_n)$ and $Y=(y_1, y_2, y_3,...,y_n)$
  - is defined as:

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

Properties of a metric distance:
- $D(X, X) = 0$
- $D(X, Y) = D(Y, X)$
- $D(X, Z) + D(Z, Y) \geq D(X, Y)$
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.

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

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:

\[
\text{precision} = \frac{|\text{Relevant} \cap \text{Retrieved}|}{|\text{Retrieved}|}
\]

\[
\text{recall} = \frac{|\text{Relevant} \cap \text{Retrieved}|}{|\text{Relevant}|}
\]

Intelligent Web Search

- F-score
  - harmonic mean of precision and recall
  - \((2 \times \text{precision} \times \text{recall}) / (\text{precision} + \text{recall})\)
  - A single consolidated measure in which recall and precision are evenly weighted.

- Combine the intelligent IR tools
  - meaning of words
  - order of words in the query
  - authority of the source

- With the unique web features
  - retrieve Hyper-link information
  - utilize Hyper-link as input
Text Mining

- Data mining in text: find something useful and surprising from a text collection;
  - text mining vs. information retrieval;
  - data mining vs. database queries.
- Document classification
  - Topic hierarchies, spam filters
- Document clustering
  - cluster documents by a common author
  - cluster documents containing information from a common source (fraud)
- Key-word based association rules

http://clusty.com
Web Structure Mining

Exploiting Hyperlink Structure
Social network analysis

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”
  - A page is important if important pages link to it
- 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. Social network analysis
3. Google
4. 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

[Graph showing in-degree distribution with power-law fit]
**Out-Links**

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.

---

**The Web Structure**

**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

**Some statistics**

- Only between 25% of the pages there is a connecting path.
- 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!
  (http://en.wikipedia.org/wiki/Small_world_phenomenon)
- Small World Graphs
  - High number of relatively small cliques
  - Small diameter
- Internet (SCC) is a small world graph
Social Network analysis

- Social network is the study of social entities (people in an organization, called actors), and their interactions and relationships.
- The interactions and relationships can be represented with a network or graph,
  - each vertex (or node) represents an actor and
  - each link represents a relationship.
- From the network, we can study the properties of its structure, and the role, position and prestige of each social actor.
- We can also find various kinds of sub-graphs, e.g., communities formed by groups of actors.

Centrality

- Important or prominent actors are those that are linked or involved with other actors extensively.

Undirected Graph: In an undirected graph, the degree centrality of an actor $i$ (denoted by $C_D(i)$) is simply the node degree (the number of edges) of the actor node, denoted by $d(i)$, normalized with the maximum degree, $n-1$.

$$C_D(i) = \frac{d(i)}{n-1}.$$ (1)

Closeness Centrality

This view of centrality is based on the closeness or distance. The basic idea is that an actor $x_i$ is central if it can easily interact with all other actors. That is, its distance to all other actors is short. Thus, we can use the shortest distance to compute this measure. Let the shortest distance from actor $i$ to actor $j$ be $d(i,j)$ (measured as the number of links in a shortest path).

Undirected Graph: The closeness centrality $C_C(i)$ of actor $i$ is defined as

$$C_C(i) = \frac{n-1}{\sum_{j \neq i} d(i,j)}.$$ (3)

The value of this measure also ranges between 0 and 1 as $n-1$ is the minimum value of the denominator, which is the sum of the shortest distances from $i$ to all other actors. Note that this equation is only meaningful for a connected graph.
Degree Prestige

Based on the definition of the prestige, it is clear that an actor is prestigious if it receives many in-links or nominations. Thus, the simplest measure of prestige of an actor \( i \) (denoted by \( P_P(i) \)) is its in-degree.

\[
P_P(i) = \frac{d_i(i)}{n-1},
\]

where \( d_i(i) \) is the in-degree of \( i \) (the number of in-links of \( i \)) and \( n \) is the total number of actors in the network. As in the degree centrality, dividing by \( n-1 \) standardizes the prestige value to the range from 0 and 1. The maximum prestige value is 1 when every other actor links to or chooses actor \( i \).

Bing Liu, Web Data Mining

PageRank and HITS

- The year 1998 was an eventful year for Web link analysis models. Both the PageRank and HITS algorithms were reported in that year.
- The connections between PageRank and HITS are quite striking.
- Since that eventful year, PageRank has emerged as the dominant link analysis model,
  - due to its query-independence,
  - its ability to combat spamming, and
  - Google’s huge business success.

Google – PageRank

- Intuition: PageRank can be seen as the probability that a “random surfer” visits a page

- A page is important if important pages link to it

PageRank

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

(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.
PageRank

- Let $\text{OutDegree}_i = \# \text{ out-links of page } i$
- Adjust $p_j$:

$$\text{PageRank}(j) = p + (1 - p) \sum_{i \neq j} \frac{\text{PageRank}(i)}{\text{OutDegree}(i)}$$

- 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

---

Google

- Keywords $w$ entered by user
- Select pages containing $w$ and pages which have in-links with caption $w$
  - Anchor text
    - Provides 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

---

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

Pages that link to a collection of authoritative pages on a broad topic
pages point to interesting links to authorities = relevant pages

Authorities

Relevant pages of the highest quality on a broad topic

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:
    $a(p) = \sum_{q \in h(p)} h(q)$
    $h(p) = \sum_{q \in a(p)} a(q)$
  - After a fixed number of iterations, pages with highest hub/authority weights form core of community

Strengths and weaknesses of HITS

- **Strength**: its ability to rank pages according to the query topic, which may be able to provide more relevant authority and hub pages.
- **Weaknesses**:
  - It is easily spammed. It is in fact quite easy to influence HITS since adding out-links in one’s own page is so easy.
  - Topic drift. Many pages in the expanded set may not be on topic.
  - Inefficiency at query time: The query time evaluation is slow. Collecting the root set, expanding it and performing eigenvector computation are all expensive operations
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 behaviours 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 regards 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.

Web Usage Mining Process

- **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
Analog – Web Log File Analyser
http://www.analog.cx/

- 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 you
  - what is being downloaded

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

Data Mining on Web Transactions

- Association Rules:
  - discovers similarity among sets of items across transactions
    \[ X \Rightarrow Y \]
  - where \( X, Y \) are sets of items, \( \alpha = \text{confidence} = P(X \cup Y) \), \( \sigma = \text{support} = P(X \cap Y) \)
- Examples:
  - 60% of clients who accessed /products/, also accessed /products/software/webminer.htm.
  - 30% of clients who accessed /special-offer.html, placed an online order in /products/software/.
  - (Actual Example from IBM official Olympics Site)
    - Badminton, Diving \( \Rightarrow \) Table Tennis \( (\alpha = 69.7\%, \ \sigma = 0.35\%) \)

Other Data Mining Techniques

- Sequential Patterns:
  - 30% of clients who visited /products/software/, had done a search in Yahoo using the keyword “software” before their visit
  - 60% of clients who placed an online order for WEBMINER, placed another online order for software within 15 days
- Clustering and Classification
  - clients who often access /products/software/webminer.html tend to be from educational institutions.
  - clients who placed an online order for software tend to be students in the 20-25 age group and live in the United States.
  - 75% of clients who download software from /products/software/demos/ visit between 7:00 and 11:00 pm on weekends.
Path and Usage Pattern Discovery

- Types of Path/Usage Information
  - Most Frequent paths traversed by users
  - Entry and Exit Points
  - Distribution of user session duration

- Examples:
  - 60% of clients who accessed /home/products/file1.html, followed the path /home ==> /home/whatsnew ==> /home/products ==> /home/products/file1.html
  - (Olympics Web site) 30% of clients who accessed sport specific pages started from the Sneakpeek page.
  - 65% of clients left the site after 4 or less references.

A Final Question

- Google Tools: search engine, Google news, Google maps, Google voice, Gmail, Calendar, Google docs, YouTube, Orkut, Google Desktop, Chrome, Android...

- Do You Trust Google to Resist Data Mining Across Services?


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

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 !!!