What is Web Mining?

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services.

There are three general classes of information that can be discovered by web mining:

- Web activity, from server logs and Web browser activity tracking.
- Web graph, from links between pages, people and other data.
- Web content, for the data found on Web pages and inside of documents.

At Scale Unlimited we focus on the last one – extracting value from web pages and other documents found on the web.

Note that there's no explicit reference to "search" in the above description. While search is the biggest web miner by far, and generates the most revenue, there are many other valuable end uses for web mining results. A partial list includes:

- Business intelligence
- Competitive intelligence
- Pricing analysis
- Events
- Product data
- Popularity
- Reputation

Four Steps in Content Web Mining

When extracting Web content information using web mining, there are four typical steps.

1. Collect – fetch the content from the Web
2. Parse – extract usable data from formatted data (HTML, PDF, etc)
3. Analyze – tokenize, rate, classify, cluster, filter, sort, etc.
4. Produce – turn the results of analysis into something useful (report, search index, etc)

Web Mining versus Data Mining

When comparing web mining with traditional data mining, there are three main differences to consider:

1. **Scale** – In traditional data mining, processing 1 million records from a database would be large job. In web mining, even 10 million pages wouldn’t be a big number.
2. **Access** – When doing data mining of corporate information, the data is private and often requires access rights to read. For web mining, the data is public and rarely requires access rights. But web mining has additional constraints, due to the implicit agreement with webmasters regarding automated (non-user) access to this data. This implicit agreement is that a webmaster allows crawlers access to useful data on the website, and in return the crawler (a) promises not to overload the site, and (b) has the potential to drive more traffic to the website once the search index is published. With web mining, there often is no such index, which means the crawler has to be extra careful/polite during the crawling process, to avoid causing any problems for the webmaster.
3. **Structure** – A traditional data mining task gets information from a database, which provides some level of explicit structure. A typical web mining task is processing unstructured or semi-structured data from web pages. Even when the underlying information for web pages comes from a database, this often is obscured by HTML markup.

Note that by “traditional” data mining we mean the type of analysis supported by most vendor tools, which assumes you’re processing table-oriented data that typically comes from a database.

Text mining

Text databases consist of huge collection of documents. They collect these information from several sources such as news articles, books, digital libraries, e-mail messages, web pages, etc. Due to increase in the amount of information, the text databases are growing rapidly. In many of the text databases, the data is semi-structured.

For example, a document may contain a few structured fields, such as title, author, publishing_date, etc. But along with the structure data, the document also contains unstructured text components, such as abstract and contents. Without knowing what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the data. Users require tools to compare the documents and rank their importance and relevance. Therefore, text mining has become popular and an essential theme in data mining.

Information Retrieval

Information retrieval deals with the retrieval of information from a large number of text-based documents. Some of the database systems are not usually present in information retrieval systems because both handle different kinds of data. Examples of information retrieval system include −

- Online Library catalogue system
- Online Document Management Systems
- Web Search Systems etc.

Note − The main problem in an information retrieval system is to locate relevant documents in a document collection based on a user's query. This kind of user's query consists of some keywords describing an information need.

In such search problems, the user takes an initiative to pull relevant information out from a collection. This is appropriate when the user has ad-hoc information need, i.e., a short-term need. But if the user has a long-term information need, then the retrieval system can also take an initiative to push any newly arrived information item to the user.

This kind of access to information is called Information Filtering. And the corresponding systems are known as Filtering Systems or Recommender Systems.

Basic Measures for Text Retrieval

We need to check the accuracy of a system when it retrieves a number of documents on the basis of user's input. Let the set of documents relevant to a query be denoted as {Relevant} and the set of retrieved document as {Retrieved}. The set of documents that are relevant and retrieved can be denoted as {Relevant} ∩ {Retrieved}. This can be shown in the form of a Venn diagram as follows −

There are three fundamental measures for assessing the quality of text retrieval −

- Precision
- Recall
- F-score

Precision

Precision is the percentage of retrieved documents that are in fact relevant to the query. Precision can be defined as −

$$\text{Precision} = |\{\text{Relevant}\} \cap \{\text{Retrieved}\}| \;/\; |\{\text{Retrieved}\}|$$

Recall

Recall is the percentage of documents that are relevant to the query and were in fact retrieved. Recall is defined as −

$$\text{Recall} = |\{\text{Relevant}\} \cap \{\text{Retrieved}\}| \;/\; |\{\text{Relevant}\}|$$

F-score

F-score is the commonly used trade-off. The information retrieval system often needs to trade-off for precision or vice versa. F-score is defined as harmonic mean of recall or precision as follows −

$$\text{F-score} = \text{recall} \times \text{precision} \;/\; (\text{recall} + \text{precision}) \;/\; 2$$