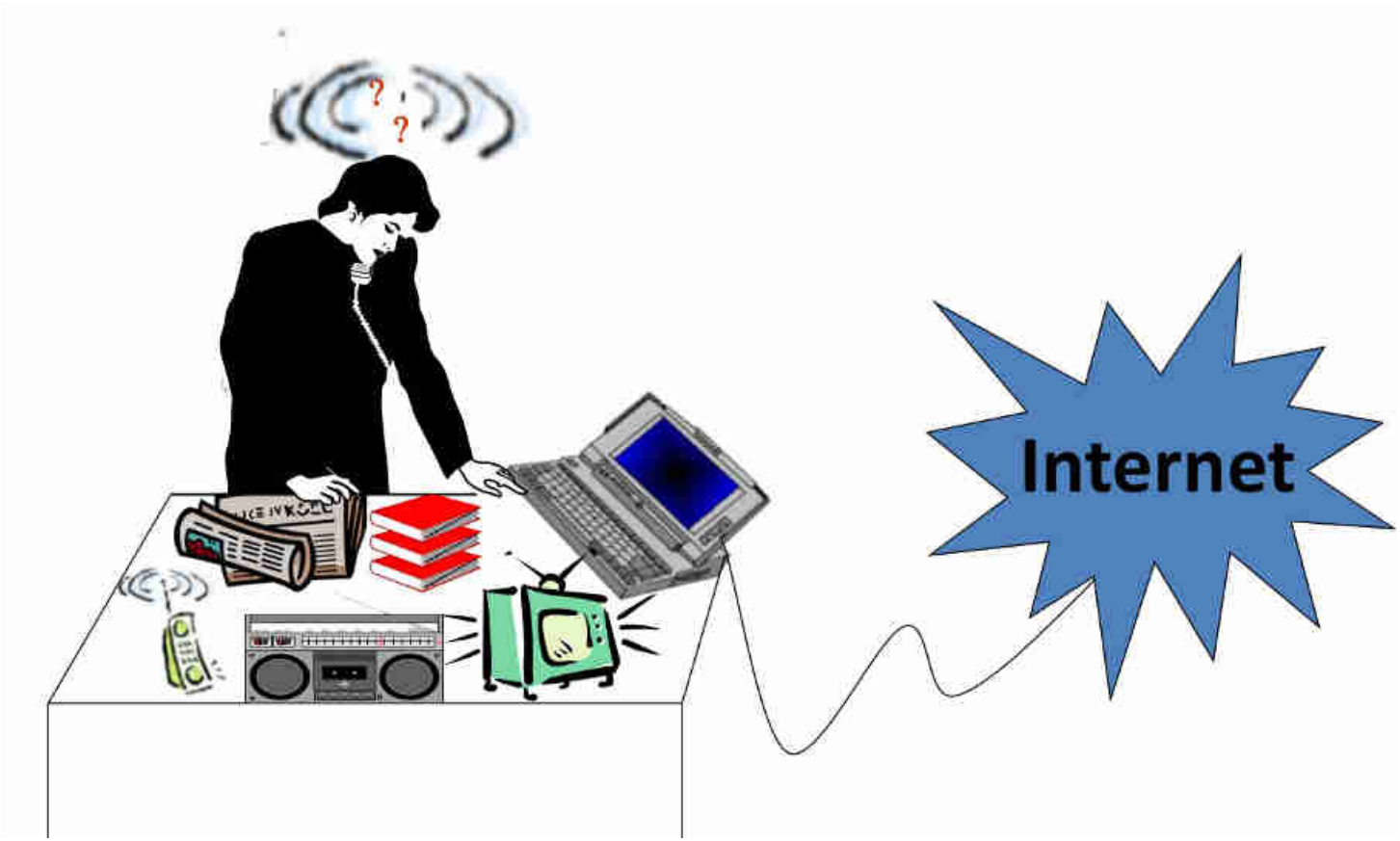
Unit 9

Advanced Concepts in Data Mining

Text Mining



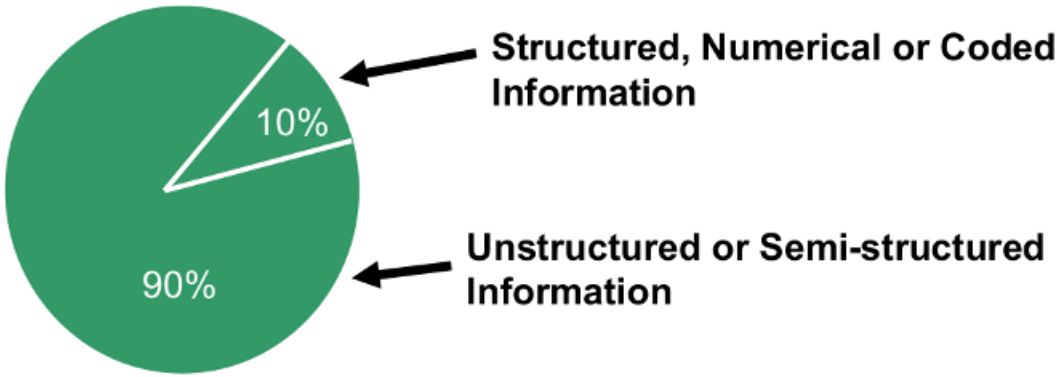
Text Mining

- Text mining is the procedure of synthesizing information, by analyzing relations, patterns, and rules among textual data.
- These procedures contains text summarization, text categorization, and text clustering.

1. **Text summarization** is the procedure to extract its partial content reflecting its whole contents automatically.
2. **Text categorization** is the procedure of assigning a category to the text among categories predefined by users
3. **Text clustering** is the procedure of segmenting texts into several clusters, depending on the substantial relevance.

Motivation for Text Mining

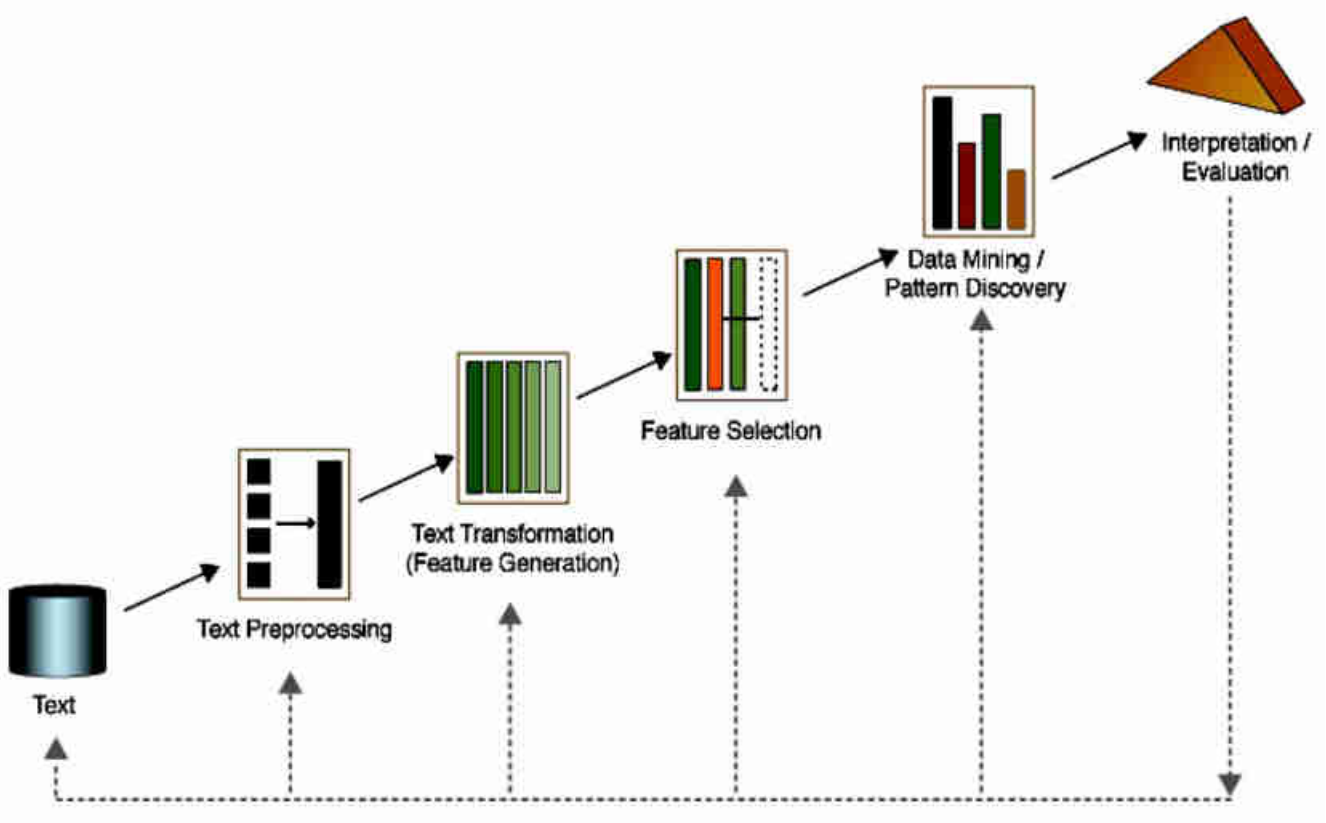
- Approximately 90% of the world's data is held in unstructured formats (Source: Oracle Corporation)
- Information intensive business processes demand that we overstep from simple document retrieval to "knowledge" discovery.



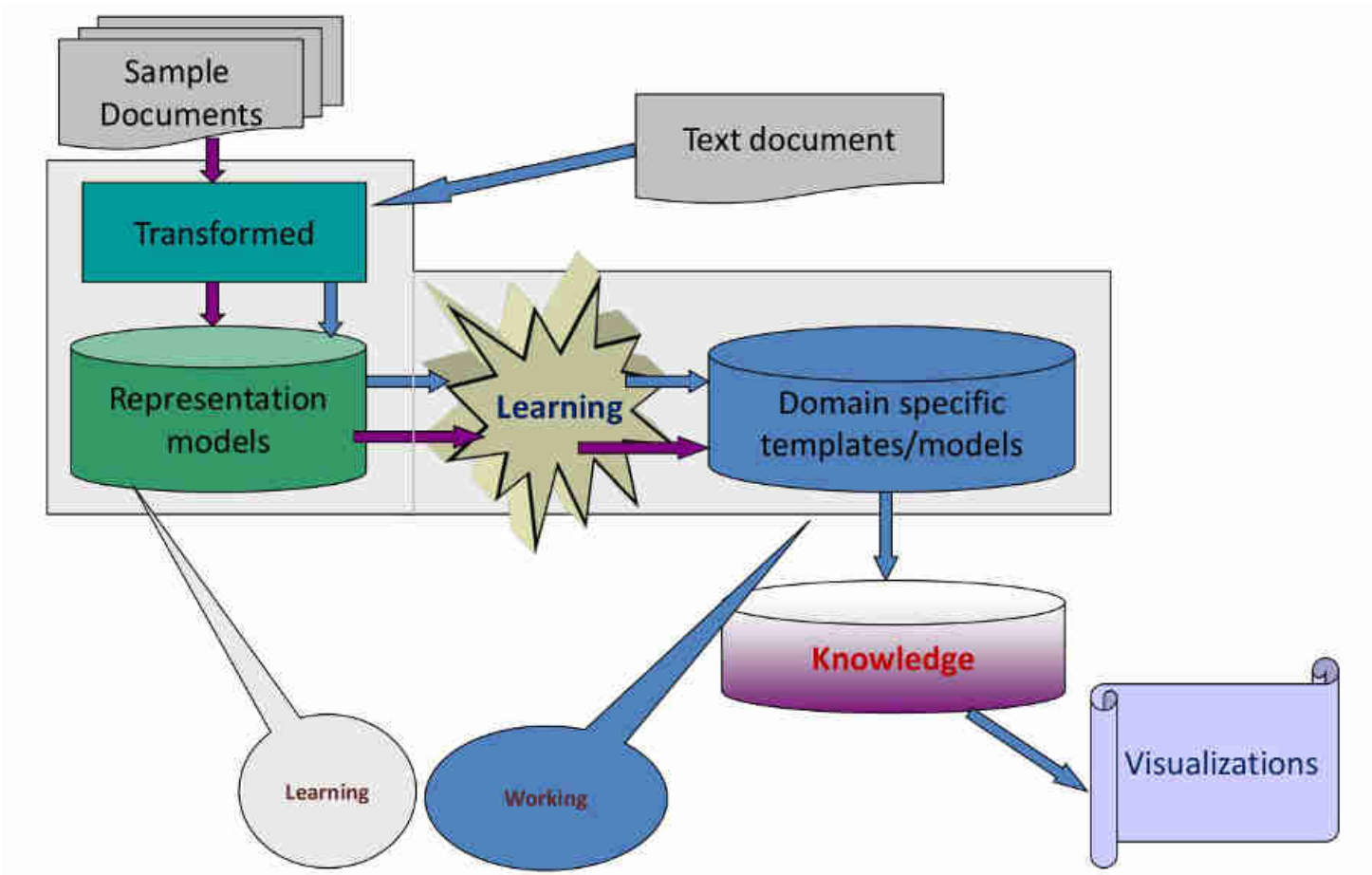
Motivation for Text Mining

- Text mining is well motivated, due to the fact that much of the world's data can be found in free text form (newspaper articles, emails, literature, etc.). There is a lot of information available to mine.
- While mining free text has the same goals as data mining, in general, extracting useful knowledge/stats/trends
- Machines can reason with relational data well since schemas are explicitly available. Free text, however, encodes all semantic information within natural language.
- Our text mining algorithms, then, must make some sense out of this natural language representation. Humans are great at doing this, but this has proved to be a problem for machines.

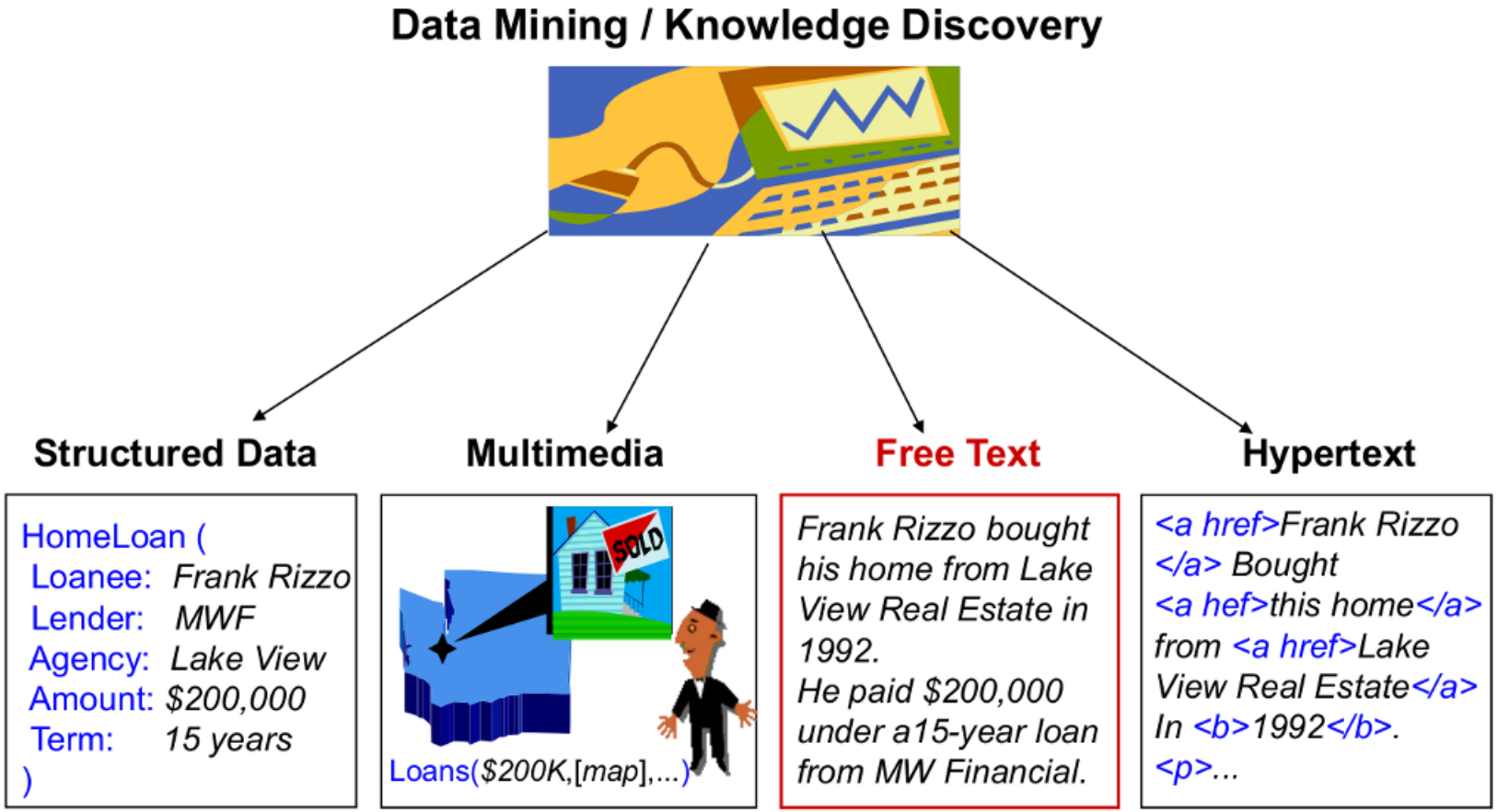
Text Mining Process



What's Text Mining



Mining Text Data: An Introduction



Text Representation Issues

- Each word has a dictionary meaning, or meanings
  - Run – (1) the verb. (2) the noun, in cricket
  - Cricket – (1) The game. (2) The insect.
  - Apple (the company) or apple (the fruit)
- Ambiguity and context sensitivity – Each word is used in various “senses”
  - I saw a man with a telescope
- Capturing the “meaning” of sentences is an important issue as well.
- Order of words in the query
  - hot dog stand in the amusement park
  - hot amusement stand in the dog park

Text Databases and IR

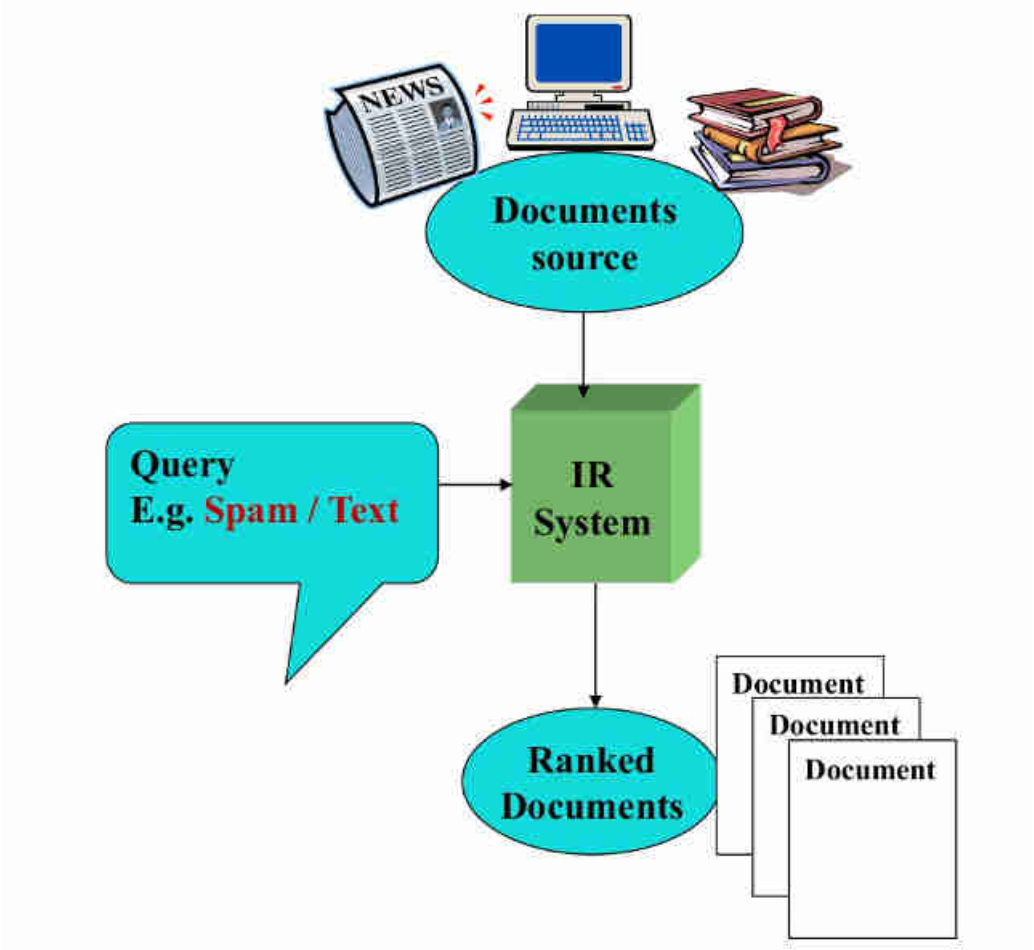
Text databases (document databases)

- Large collections of documents from various sources: news articles, research papers, books, digital libraries, e-mail messages, and Web pages, library database, etc.
- Data stored is usually semi-structured
- Traditional information retrieval techniques become inadequate for the increasingly vast amounts of text data

Information retrieval

- A field developed in parallel with database systems Information is organized into (a large number of) documents
- Information retrieval problem: locating relevant documents based on user input, such as keywords or example documents

Information Retrieval



Basic Measure for Text Retrieval

Precision: This is the percentage of retrieved documents that are in fact relevant to the query (i.e., "correct" responses). It is formally defined as

$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|}.$$

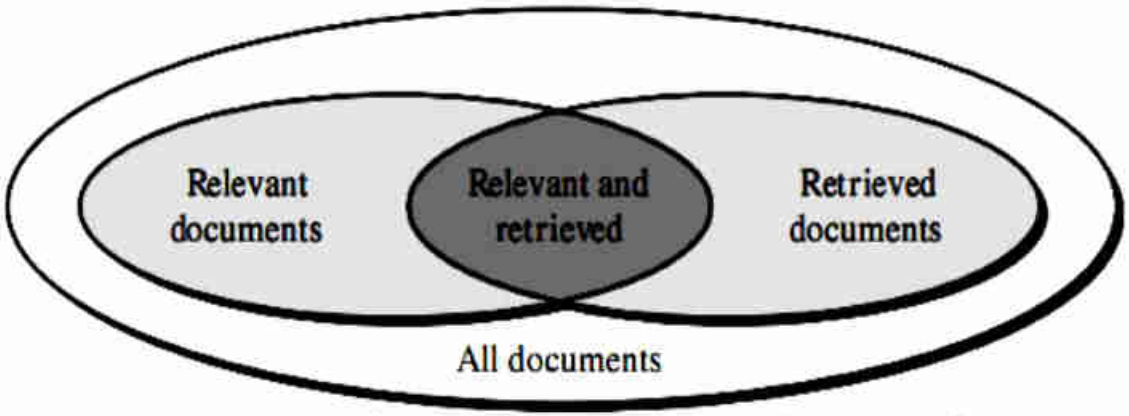
Recall: This is the percentage of documents that are relevant to the query and were, in fact, retrieved. It is formally defined as

$$recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}.$$

Basic Measure for Text Retrieval

F-Score: An information retrieval system often needs to trade off recall for precision or vice versa. One commonly used trade-off is the F-score, which is defined as the harmonic mean of recall and precision:

$$F\_score = \frac{recall \times precision}{(recall + precision)/2}.$$



Relationship between the set of relevant documents and the set of retrieved documents

Application of Text Mining

Text mining system provides a competitive edge for a company to process and take advantage of a large quantity of textual information.
The potential applications are countless. We highlight a few below.

- **Customer profile analysis**, e.g., mining incoming emails for customers' complaint and feedback.
- **Patent analysis**, e.g., analyzing patent databases for major technology players, trends, and opportunities.
- **Information dissemination**, e.g., organizing and summarizing trade news and reports for personalized information services.
- **Company resource planning**, e.g., mining a company's reports and correspondences for activities, status, and problems reported.

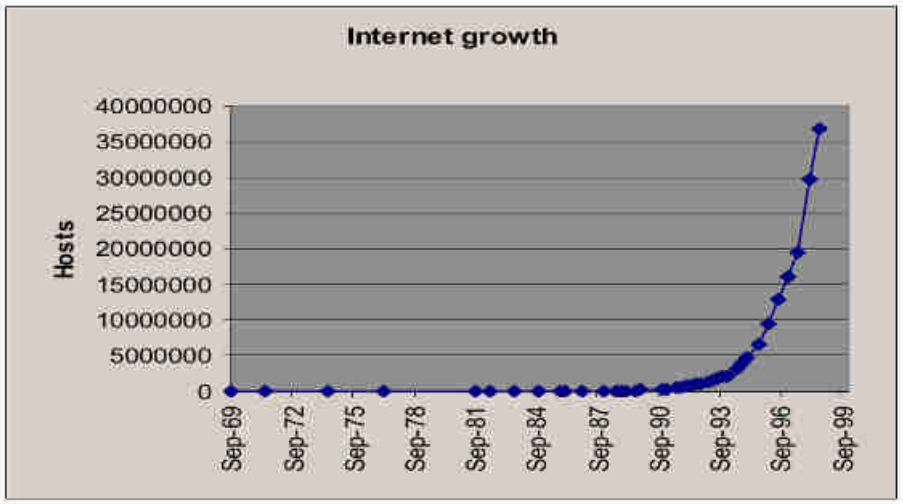
Assignment

- Text Mining Vs Data Mining
- Intelligent Miner for Text(IMT)

Mining World Wide Web (WWW)

- was coined by *Orem Etzioni (1996)* to denote the use of data mining techniques to automatically discover Web documents and services, extract information from Web resources, and uncover general patterns on the Web.

Why Mining the World-Wide Web

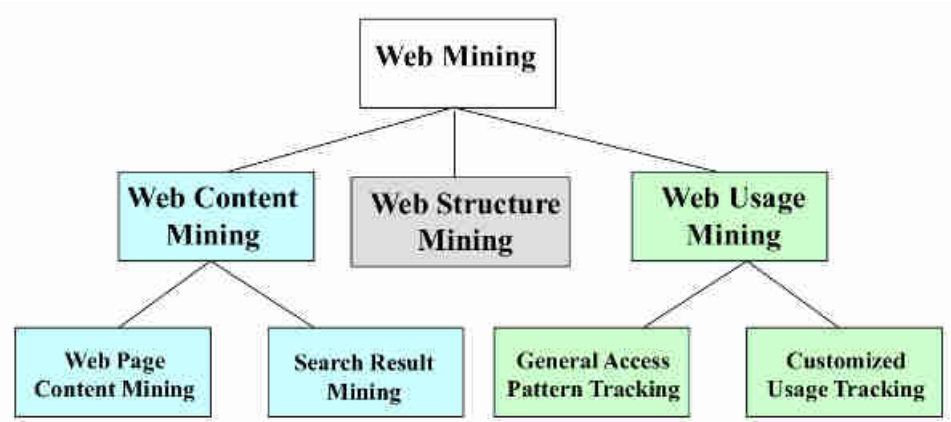


- Growing and changing very rapidly
- Broad diversity of user communities
- Only a small portion of the information on the Web is truly relevant or useful
  - 99% of the Web information is useless to 99% of Web users
  - How can we find high-quality Web pages on a specified topic?

Web Mining: A More Challenging Task

- Searches for
  - Web access patterns
  - Web structures
  - Regularity and dynamics of Web contents
- Problems
  - The “abundance” problem
  - Limited coverage of the Web: hidden Web sources, majority of data in DBMS
  - Limited query interface based on keyword-oriented search
  - Limited customization to individual users

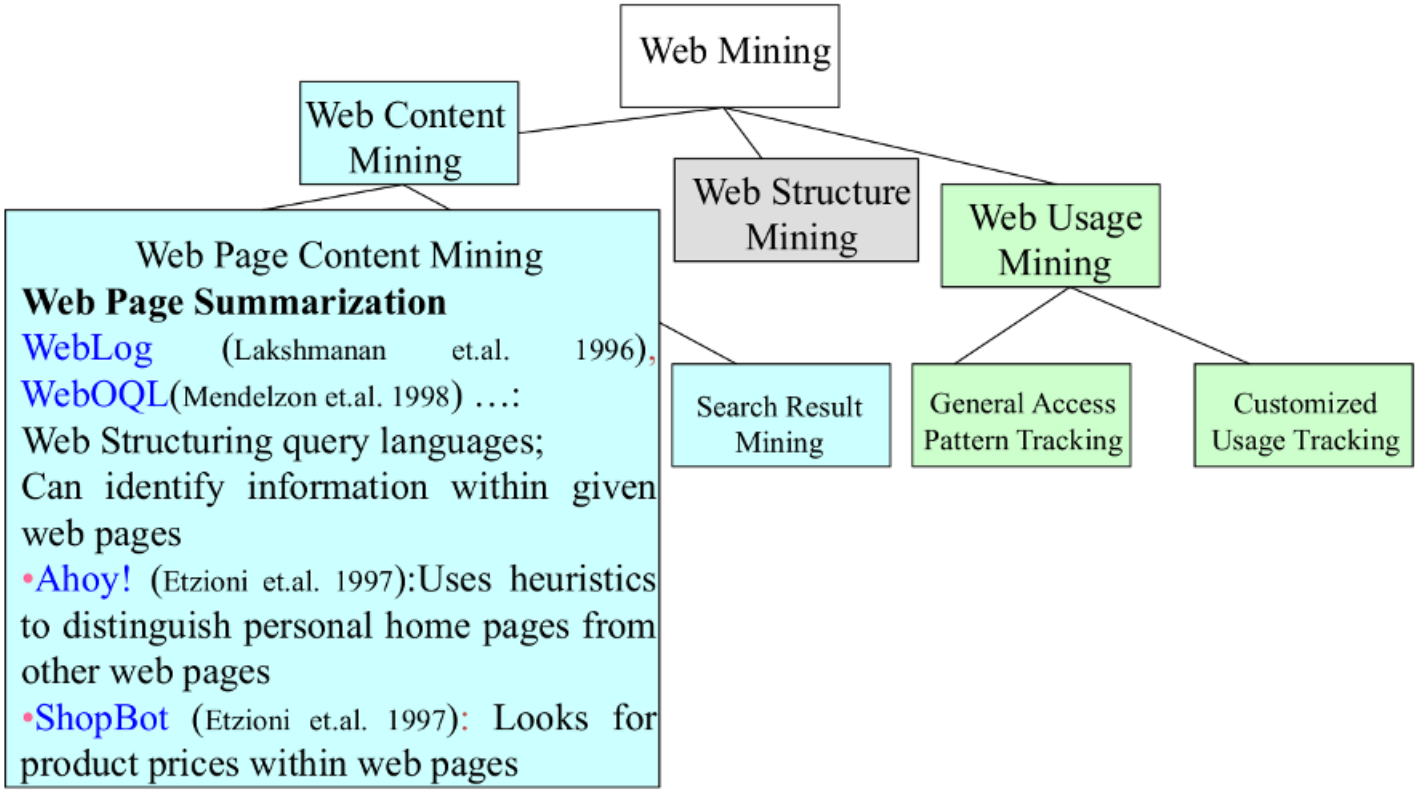
Web Mining Taxonomy



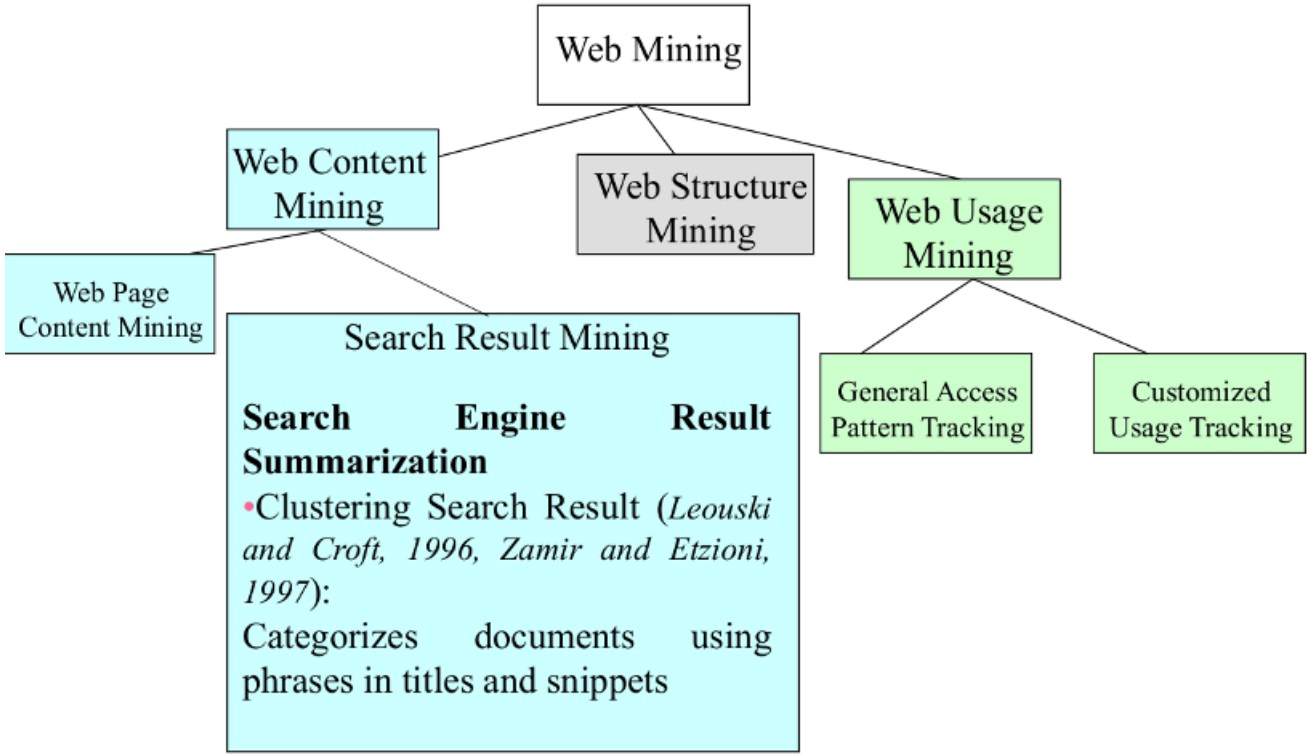
Web Mining research can be classified into three categories:

- **Web content mining** refers to the discovery of useful information from Web contents, including text, images, audio, video, etc.
- **Web structure mining** studies the model underlying the link structures of the Web. It has been used for search engine result ranking and other Web applications.
- **Web usage mining focuses** on using data mining techniques to analyze search logs to find interesting patterns. One of the main applications of Web usage mining is its use to learn user profiles.

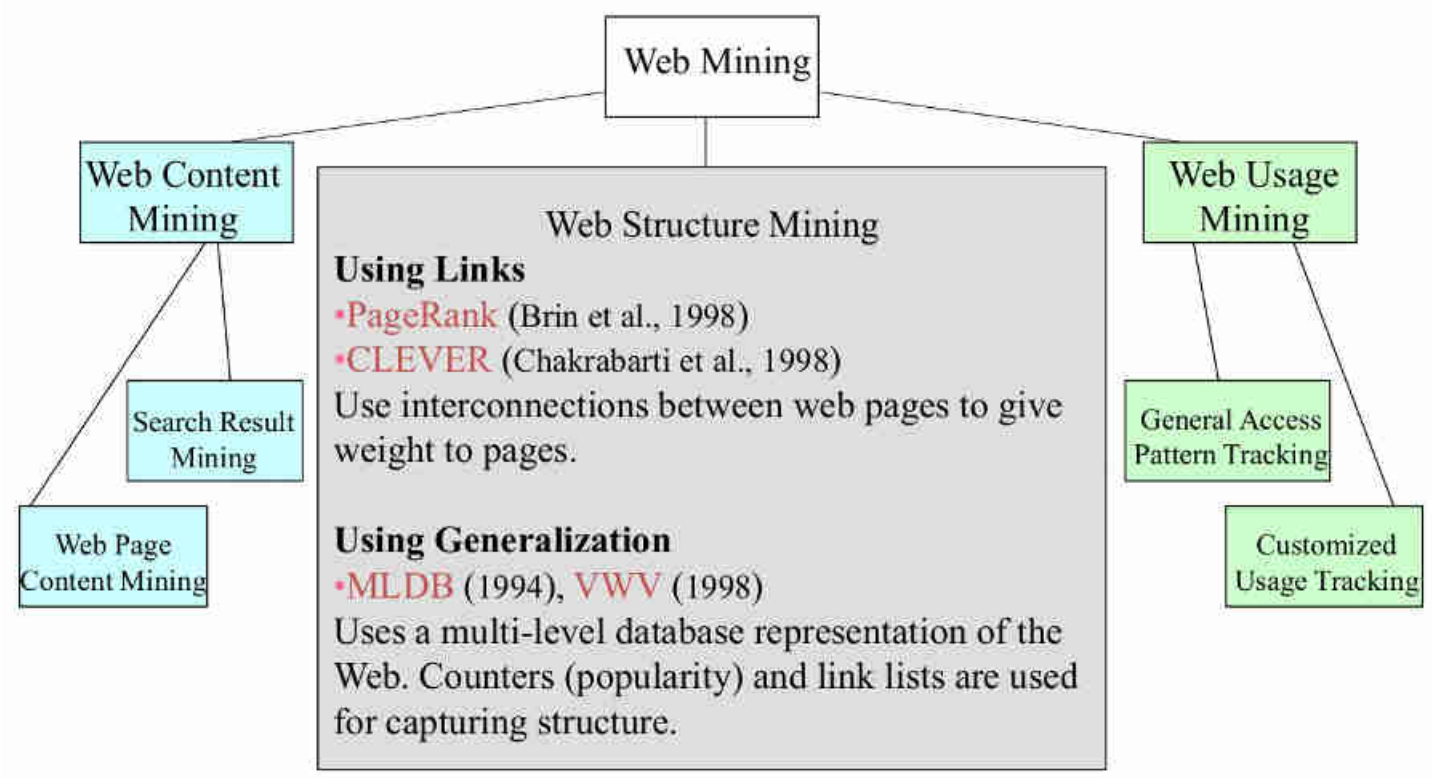
Mining the World-Wide Web



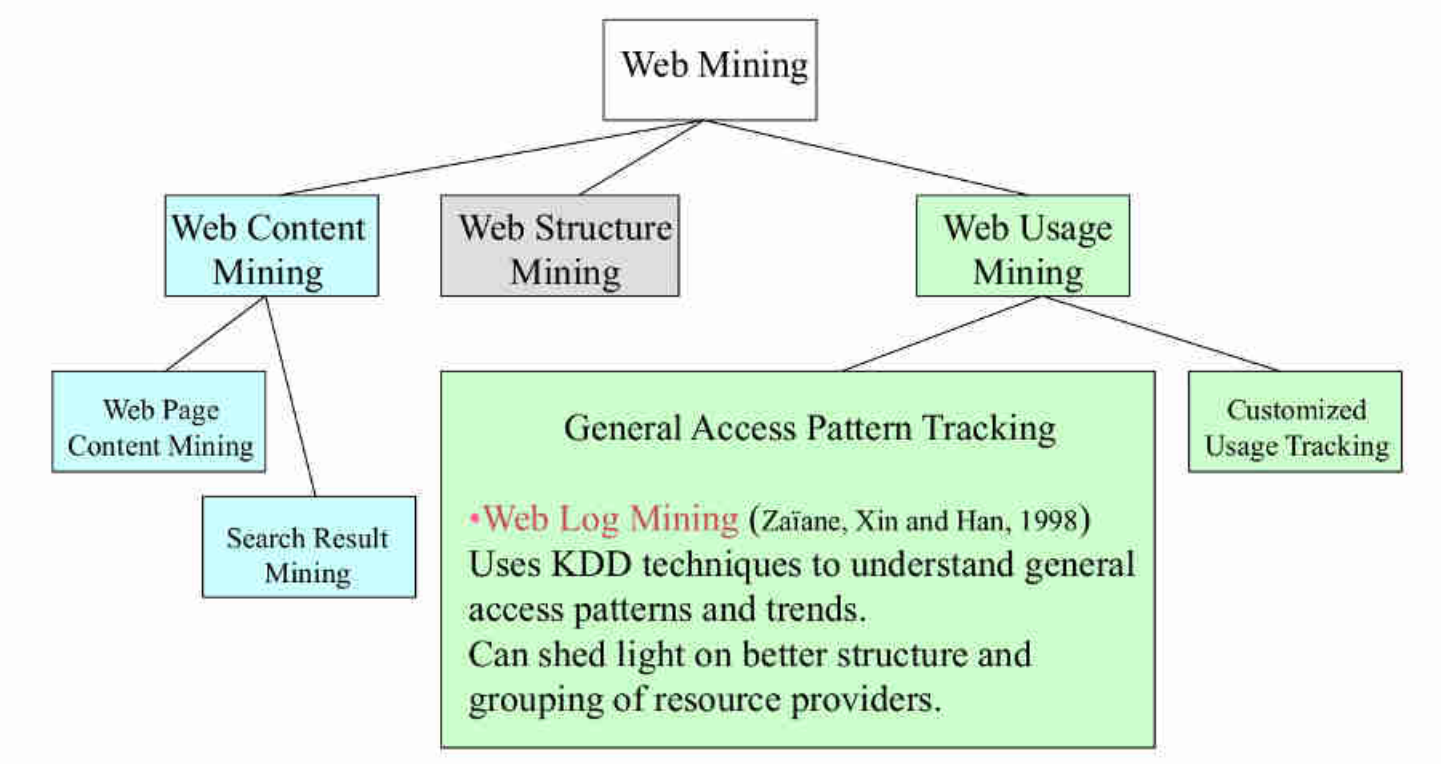
Mining the World-Wide Web



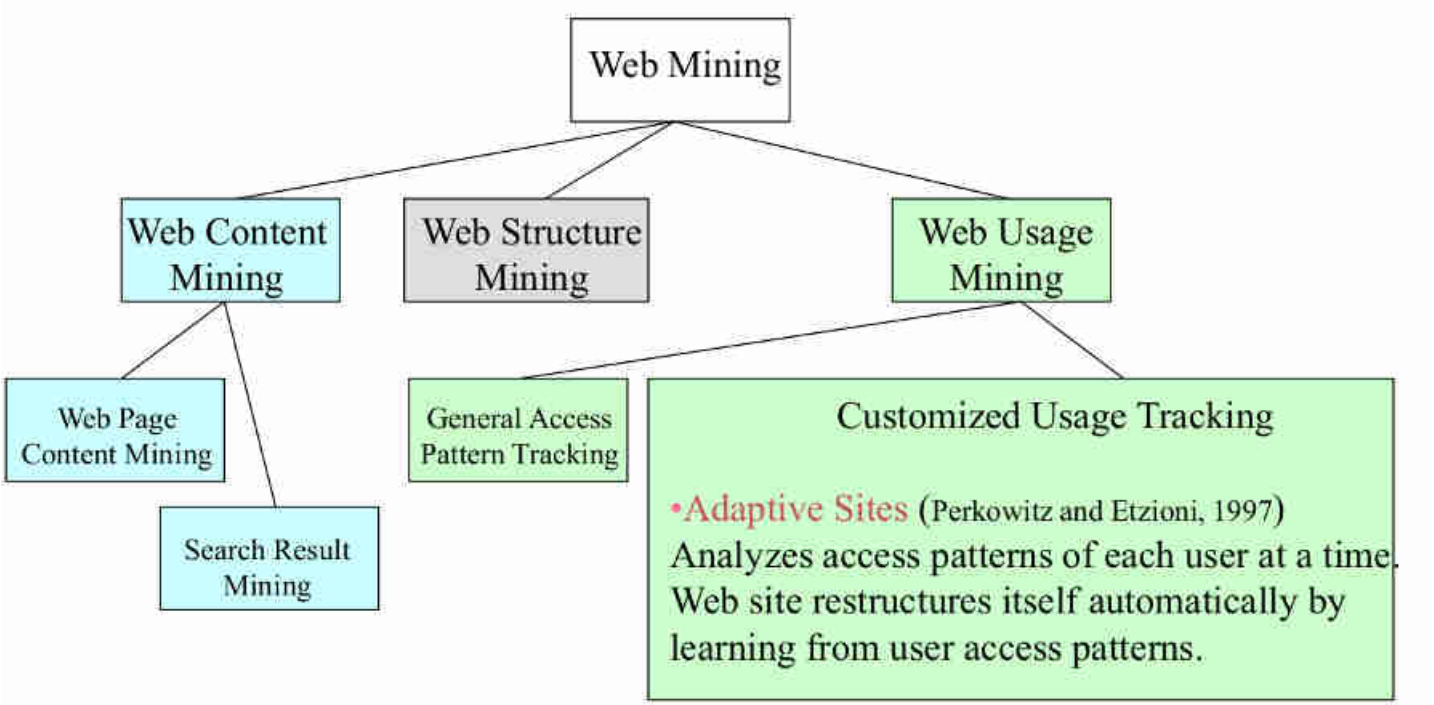
Mining the World-Wide Web



Mining the World-Wide Web



Mining the World-Wide Web



Web Content Mining

- The extraction of certain information from the unstructured raw data text of unknown structures is referred to as Web content mining.
- A set of information extraction tools is brought forward in order to identify and collect content items, such as Text Extraction and Wrapper Induction.
- It is the process of extraction of information from web document, video, audio, text, structured records such as lists and tables

Web Structure Mining

- Web structure mining (WSM) usually operates on the hyperlink structure of web pages.
- WSM focuses on sets of pages, ranging from a single website to the Web as a whole.
- WSM exploits the additional information that is contained in hypertext.

Web structure mining algorithms

The **PageRank algorithm** is computed by weighting each in-link to a page proportionally to the quality of the page containing the in-link (Brin & Page, 1998)

The qualities of these referring pages also are determined by PageRank. Thus, a page p is calculated recursively as follows

$$PageRank(p) = (1-d) + d \times \sum_{\substack{all\ q\ linking \\ to\ p}} \left(\frac{PageRank(q)}{c(q)} \right)$$

where *d is a damping factor between 0 and 1,*
c(q) is the number of out-going links in a page q.

Web structure mining algorithms

Kleinberg (1998) proposed the HITS **(Hyperlink-Induced Topic Search)** algorithm, which is similar to PageRank.

- **Authority pages:** high-quality pages related to a particular search query.
- **Hub pages:** pages provide pointers to other authority pages.

A page to which many others point should be a good authority, and a page that points to many others should be a good hub.

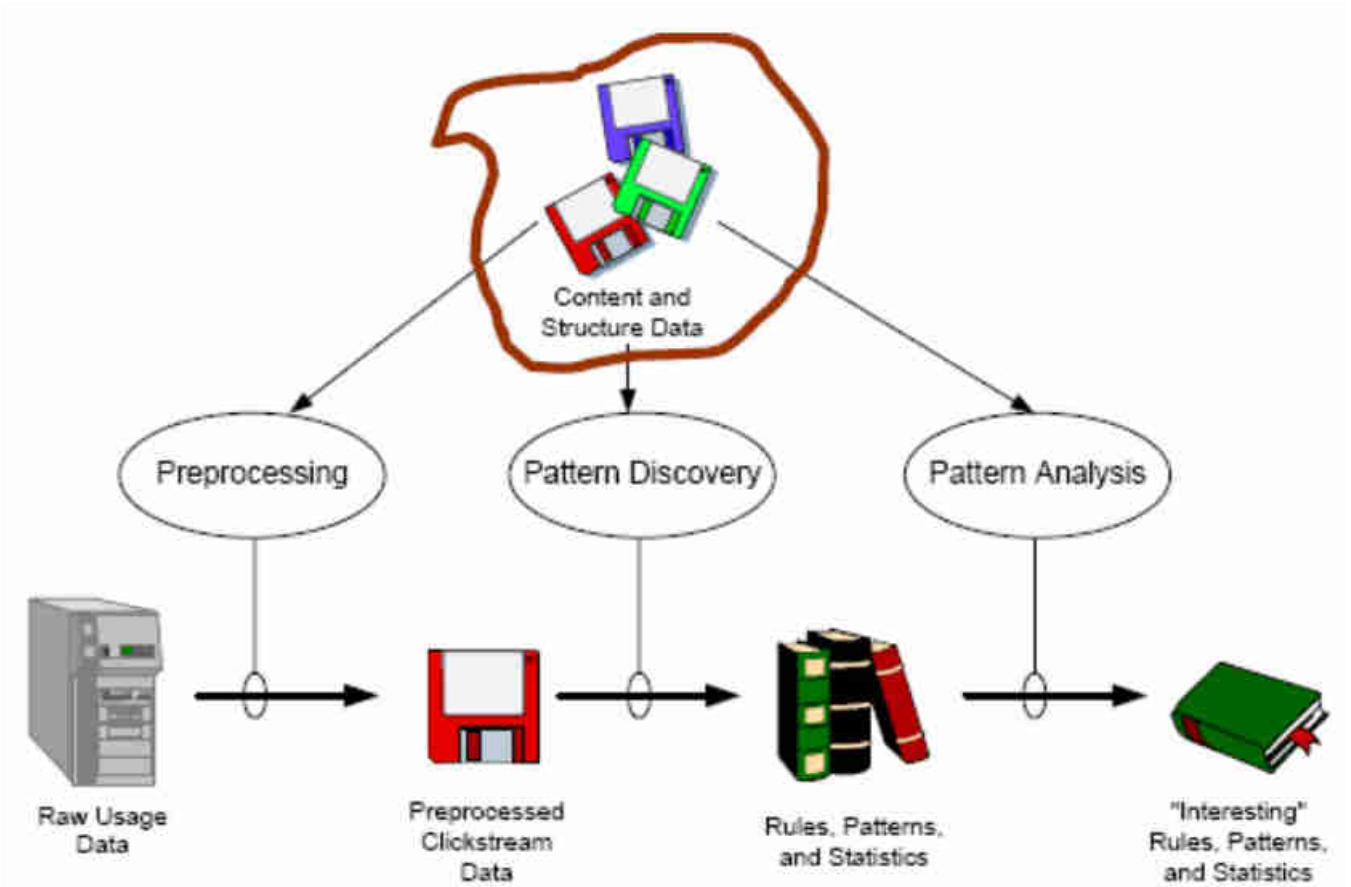
$$AuthorityScore(p) = \sum_{\text{all } q \text{ linking to } p} (HubScore(q))$$

$$HubScore(p) = \sum_{\text{all } r \text{ linking from } p} (AuthorityScore(r))$$

Web Usage Mining

- Web usage mining (WUM) focuses on records of the requests made by visitors to a website, most often collected in a web server log.
- The content and structure of web pages, and in particular those of one website, reflect the intentions of the authors and designers of the pages, and the underlying information architecture

Web Usage Mining → Procedure



Assignment

Note: This Assignment holds 2% of internal marks. Plagiarism will be checked (Among with the friends too)

- Mining complex types of data include object data, spatial data, multimedia data, time series data, text data, web data and more. How can we mine object data, multimedia data, time series data and spatial databases? Explain them individually.

- Explain the procedure for web usage mining. Explain in detail.