Web Mining

Margherita Berardi

LACAM

Dipartimento di Informatica

Università degli Studi di Bari

berardi@di.uniba.it

Bari, 24 Aprile 2003

Overview

- Introduction
- Knowledge discovery from text (Web Content Mining)
- Knowledge discovery from links (Web Structure Mining)
- Knowledge discovery from usage data (Web Usage Mining)
- Open Issues

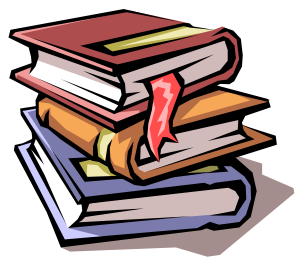
Overview

- Introduction
  - ✓ Information overload
  - ✓ The need for intelligent information access
  - ✓ Knowledge discovery approaches
    - What is Data Mining?
    - Knowledge Management and Web Mining: which relationship?
- Knowledge discovery from text (Web Content Mining)
- Knowledge discovery from links (Web Structure Mining)
- Knowledge discovery from usage data (Web Usage Mining)
- Open Issues

The Web and the information overload problem

The Web continuously introduces new capabilities and attractes many people

More and more people start to use it...



...the amount of online information increases

...and users have to face up the overload of information

The Web

- ✓ Over 300 billion pages online, ~ 4500 terabytes
- ✓ More than 100 million of queries per day
- ✓ More than 600 million of users online
- ✓ More than 9 million Web sites
- ✓ Highly Dynamic
  - 1 million new pages each day
  - A page changes in few weeks
  - Less than 50% of Web sites will be there next year
- ✓ *"99% of online information is of no interest to 99% of the people"* (The Abundance Problem)

Information access services...

- A lot of services aim to help the user to access to online information and products:
  - ✓ YAHOO!
  - ✓ altavista
  - ✓ AutoTrader
  - ✓ Lycos
  - ✓ excite
  - ✓ amazon.com
  - ✓ ...

...which limitations?

- ✓ 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.
- ✓ 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 of individual users.

The need for intelligent information access

What is Web Mining?

Web Mining is the use of data mining techniques to automatically discover and extract information from the Web. It could be successfully used to solve information overload probelm.

Web Mining

- Four different activities:
  - ✓ Resource discovery: locating documents and services in the Web.
  - ✓ Information selection and pre-processing: extracting specific information from newly discovered resources.
  - ✓ Generalization: discovering general patterns at individual web sites as well as across multiple sites.
  - ✓ Analysis: validation and/or interpretation of the mined patterns

What is Data Mining?

- Data mining is the efficient discovery of previously unknown patterns in large databases.
- This may be simply extracting information in a more meaningful way (extended SQL, OLAP) to identifying patterns inherent in the data (statistics, machine learning).
- Much time and energy is spent analysing MIS reports in order to gain knowledge from this summary transactional data.
- Data mining attempts to infer knowledge directly from the data (in transaction systems) leading to more timely, and often more accurate, decision making.

WM & DM

Web Mining is an extension of Data Mining techniques to discover and organize information from the Web.

This machine learning ability allows modification of search criteria automatically before the next execution, once particular patterns or trends have been discovered in the data searched.

WM & KM

What is Knowledge Management ?

The process of accumulating and creating knowledge efficiently, managing a knowledge base, and facilitating the sharing of knowledge so that it can be applied effectively throughout the organization.

KM: which technologies?

- Data warehouses and data marts
- Intranets and Web (KM infrastructure)
- Help-desk software
- Document management and text retrieval
- Web-based search and retrieval tools
- Data mining tools
- More specifically, Web mining tools

Overview

- Introduction
- Knowledge discovery from text (Web Content Mining)
  - ✓ Text Mining
  - ✓ Information filtering and retrieval
  - ✓ Information extraction
  - ✓ Ontology learning
- Knowledge discovery from links (Web Structure Mining)
- Knowledge discovery from usage data (Web Usage Mining)
- Open Issues

Web Content Mining

- The discovery of useful information from the Web contents.
- Goals:
  - ✓ Organize documents into categories.
  - ✓ Assign new documents to the categories.
  - ✓ Retrieve information that matches a user query.
- Dominating statistical idea:
  - ✓ TFIDF=term frequency * inverse document frequency

Web Content Mining: which data?

- ✓ Web content consists of several types of data, such as textual, image, audio, video, metadata, hyperlinks.
- ✓ Different data representation and different methods for different kind of content:
  - Unstructured, such as free-text
  - Semi-structured, such as HTML and XML documents
  - Structured, such as dynamically generated tables from databases

Web Content Mining

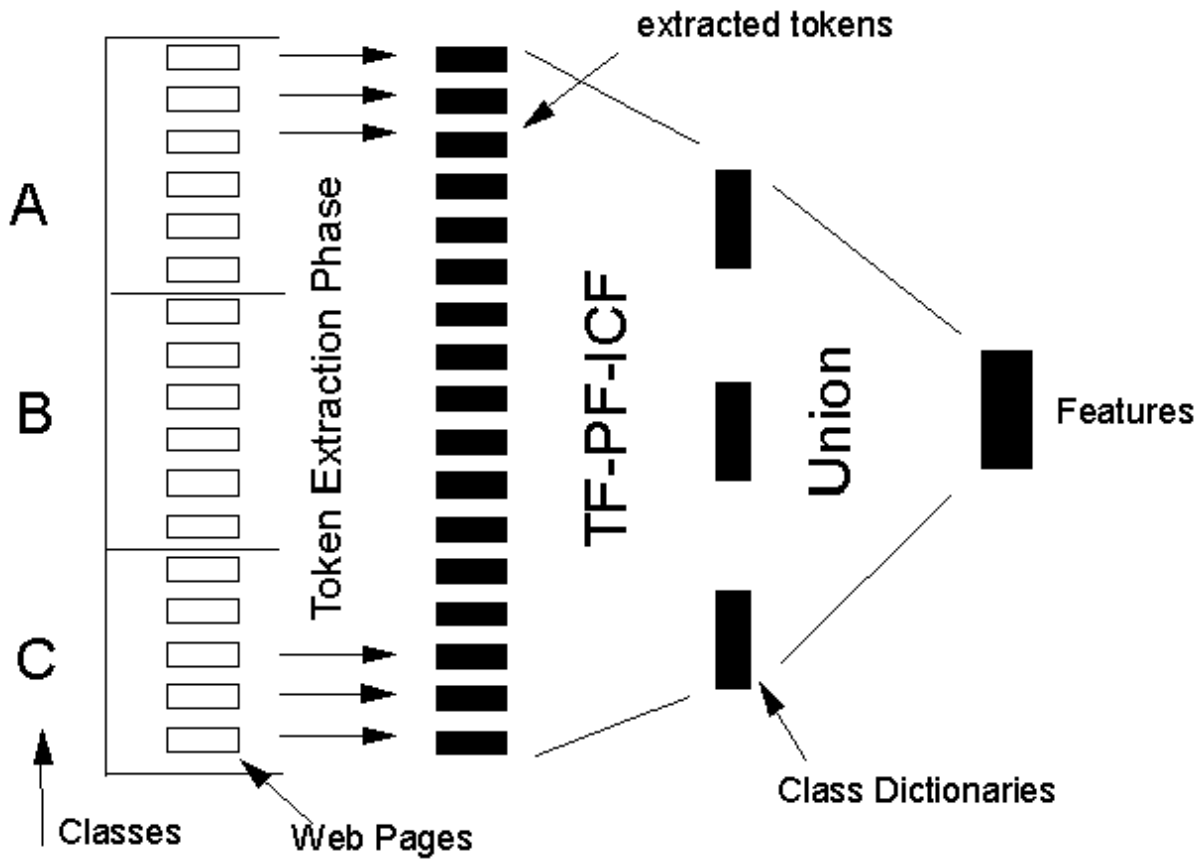
Unstructured and semi-structured text documents:

- ✓ Representation: bag-of-words, n-grams, phrase-based, term based, named entities, relational.
  - ✓ Documents are represented as feature vectors, where features correspond for example to specific words (*1-grams*) or a sentence of n words (*n-grams*).
- ✓ Patterns: classification rules, document clusters, associations rules.
- ✓ Tasks: text classification or categorization, hierarchical clustering, finding/extracting key phrases, event detection and tracking.

Web Content Mining: feature extraction and selection

- All training documents are initially tokenized.
- Some shorter tokens, HTML tags, punctuation marks, numbers and stopwords are removed.
- Removal of suffixes (-s, -es, -ies, -ed, -ing, ...)
- Stemming
- How to select discriminant tokens?

From documents to features



Some methods for a classification task

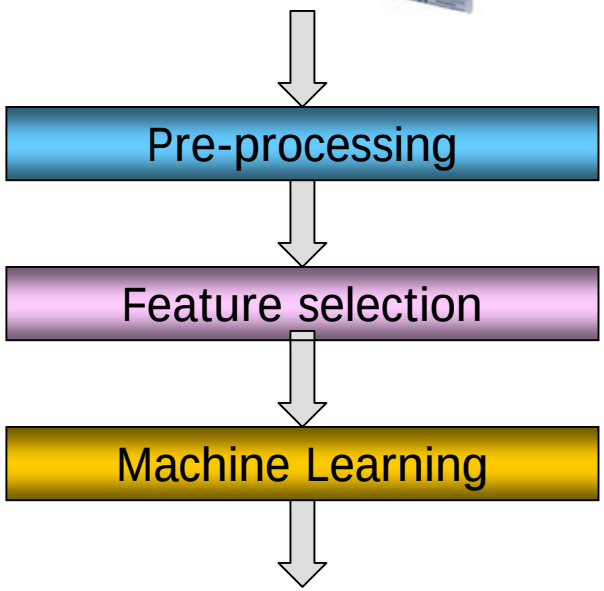
- How assign a Web page to a class ?
  1. By sequentially testing feature values according to a *decision tree*.
  2. By computing the distance from the *centroids* of the different classes.
  3. By computing the *Bayesian posterior probability*.
  4. By computing the distance from all training documents (*k-nearest-neighbor*).
- In the first three ways, a training phase is necessary. In the fourth way, training instances are used only when the system is asked to classify a new page.

Text Mining

- Knowledge discovery in textual data.
- Text is often amorphous, and difficult to deal with (e.g., email messages, open-ended comments on a questionnaire or suggestion form, patients' descriptions of their symptoms, searches of written historical records, etc.).
- Text Mining is not assigning documents to categories, but learning document classifiers.
- Text Mining is not extracting information from text, but learning information extraction patterns.

Information filtering and retrieval

Training documents (pre-classified)



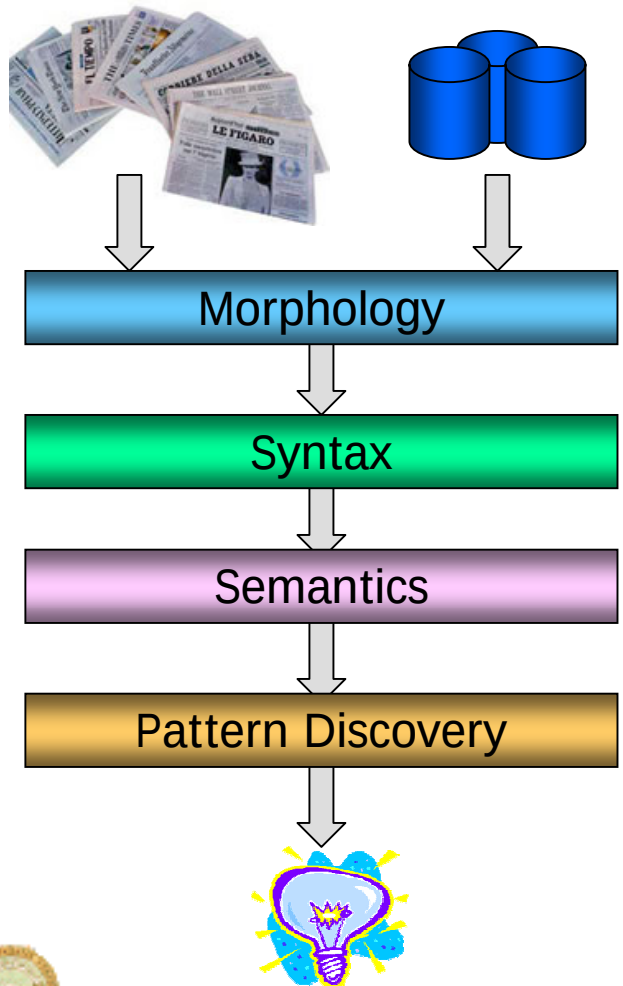
Stopword removal
Stemming
Bag-of-words coding

Statistical selection of characteristic terms

Supervised classifier learning

Category models

Information Extraction



Unstructured text & database schema

Lemmatization, sentence and word separation, Part-of-speech tagging

Shallow syntactic parsing

Named-entity recognition
Co-reference resolution
Sense disambiguation

IE pattern discovery

IE patterns

Ontology learning

Pre-processing

Feature compression

Machine Learning

Training documents (unclassified)

Stopword removal, Stemming, Syntactic/Semantic analysis, Bag-of-words coding

Hand-made thesauri (Wordnet)

Term co-occurence (LSI)

Unsupervised learning (clustering and association discovery)

Ontologies

Overview

- Introduction
- Knowledge discovery from text (Web Content Mining)
- Knowledge discovery from links (Web Structure Mining)
  - ✓ Usefulness of hyperlink information
  - ✓ How analize link structures?
- Knowledge discovery from usage data (Web Usage Mining)
- Open Issues

Web Structure Mining

- The discovery of the model underlying the link structures of the Web.
- Goal: categorization of Web pages and generation of information based on similarity among different Web sites
  - ✓ Discover authority sites (highly referenced sites on a topic)

Which useful information can be derived from the link structure of the Web?

Some statistics:

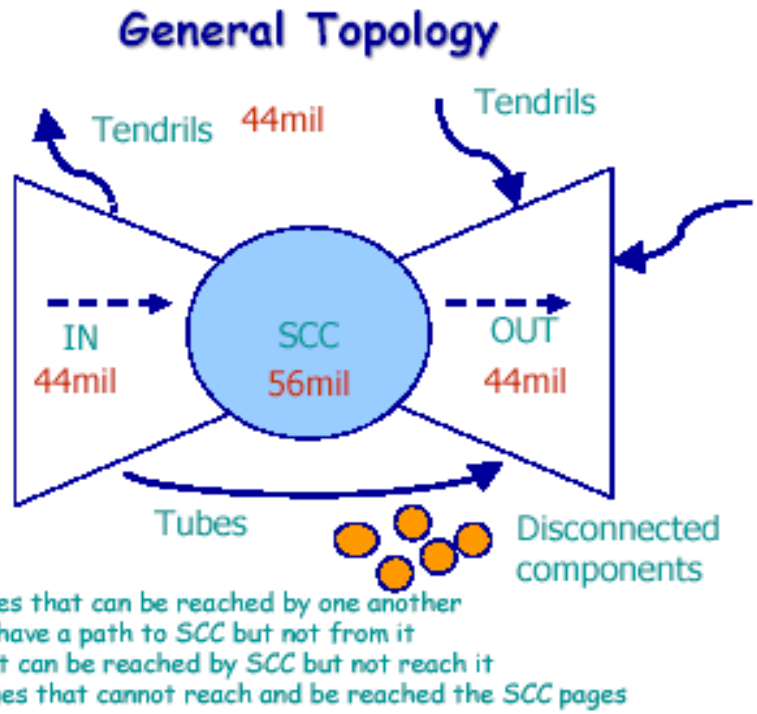
✓Only between 25% of the pages there is a connecting path

✓but if there is a path:

- directed: average length < 17
- undirected: average length < 7 (!!!)

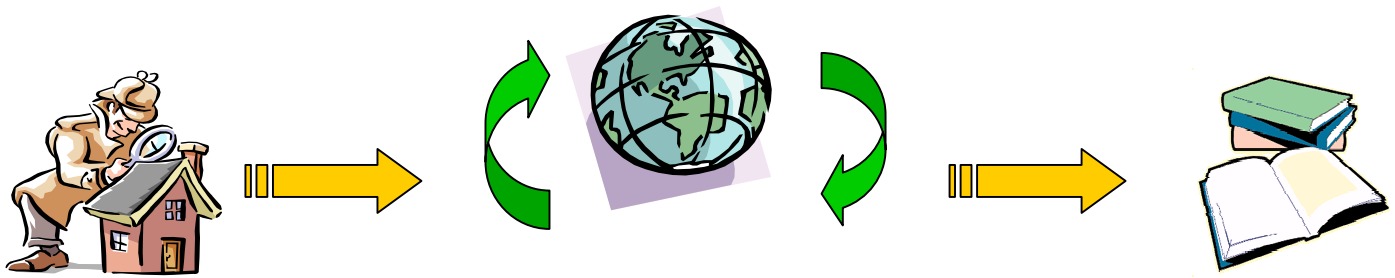
Internet (SCC)

is a small world graph



Google

- Search engine that uses link structure to calculate a quality ranking (Page Rank) for each page.
- Intuition: Page Rank can be seen as the probability that a random surfer visits a page.
- A page is important if many important pages link to it.
- Common scenario:



A user enters keywords **W**

The search engine selects pages containing **W** and pages which have in-links with caption **W**

Ranking pages according to importance

Link Structure Analysis: HITS

- The Hyperlink-Induced Topic Search uses hyperlink structure to identify authorative 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.

Link Structure Analysis: HITS

- Discovering the Hubs and Authorities on a specific topic/query involves the following steps:
  - ✓ collect a set of pages S (returned by search engine);
  - ✓ expand the base set to contain pages that point to or are pointed to by pages in the initial set;
  - ✓ iteratively update hub weight h(p) and authority weight a(p) for each page;
  - ✓ after a fixed number of iterations, return a list of the pages ranked by their hub/authority weights.

Crawling & Spidering

- Automatic navigation through the Web by robots with the aim of indexing the Web.
- Crawling v. Spidering:inter-site v. intra-site navigation.
- Underlying assumption similar to HITS: thematically similar pages are linked

Overview

- Introduction
- Knowledge discovery from text (Web Content Mining)
- Knowledge discovery from links (Web Structure Mining)
- Knowledge discovery from usage data (Web Usage Mining)
  - ✓ Web Personalization
  - ✓ Discovering generic user modeling
- Open Issues

Web Usage Mining

- The discovery of patterns in data generated by surfers behaviours:
  - Web server access logs, user queries, cookies (server level collection);
  - proxi server logs (proxi level collection);
  - Java and Javascript agents (client level collection);
  - browser logs, user profiles, registration data, user sessions or transactions, bookmark data, mouse clicks and scrolls;
  - any other results of interacions ...

Web Usage Mining

- Goal: predicting user behaviour while the user interacts with the Web.
- Two main applications:
  - ✓ learning a user profile or user modeling in adaptive interfaces,
  - ✓ learning user navigation patterns.

Discovering generic user modeling

- Constructing models that can be used to adapt the system to the user's requirements.
- Different type of requirement: interests (sports, finance, politics...), knowledge level, preferences, etc.
- Different types of model: personal – generic.

Discovering generic user modeling

- Discovering stereotypes: extract models that represent a type of user with personal characteristics.
- Discovering communities: extract models that represent a group of users with common preferences.

Web Personalization

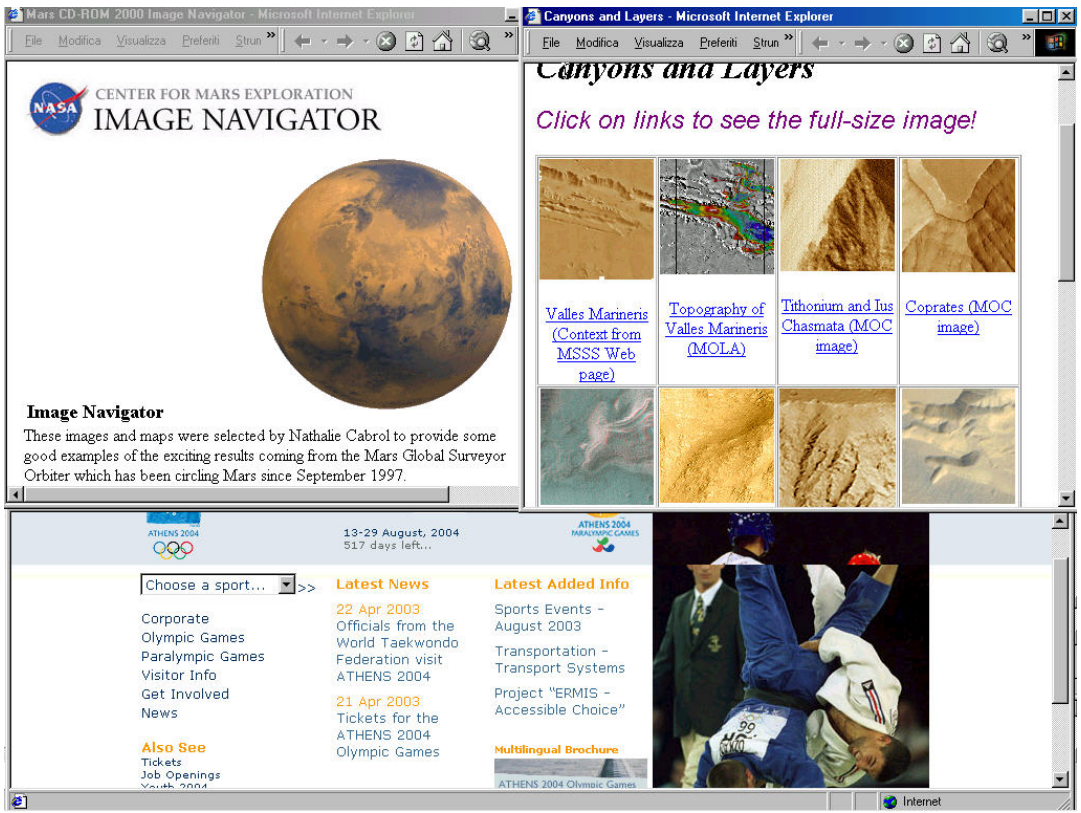
- Improve effectiveness of the information on Web sites by adapting the Web site design or by biasing the user's behaviour towards satisfying the goals of the site.
- Construction of a separate model for each user and use of this model to:
  - ✓ help focus on interesting Web sites,
  - ✓ modify the structure and content of a site,
  - ✓ adapt the Web interface.

Overview

- Introduction
- Knowledge discovery from text (Web Content Mining)
- Knowledge discovery from links (Web Structure Mining)
- Knowledge discovery from usage data (Web Usage Mining)
- Open Issues
  - Multimedia content
  - Discovery in the Web graph
  - Dealing with unexplored sizes

Dealing with Multimedia Web Data

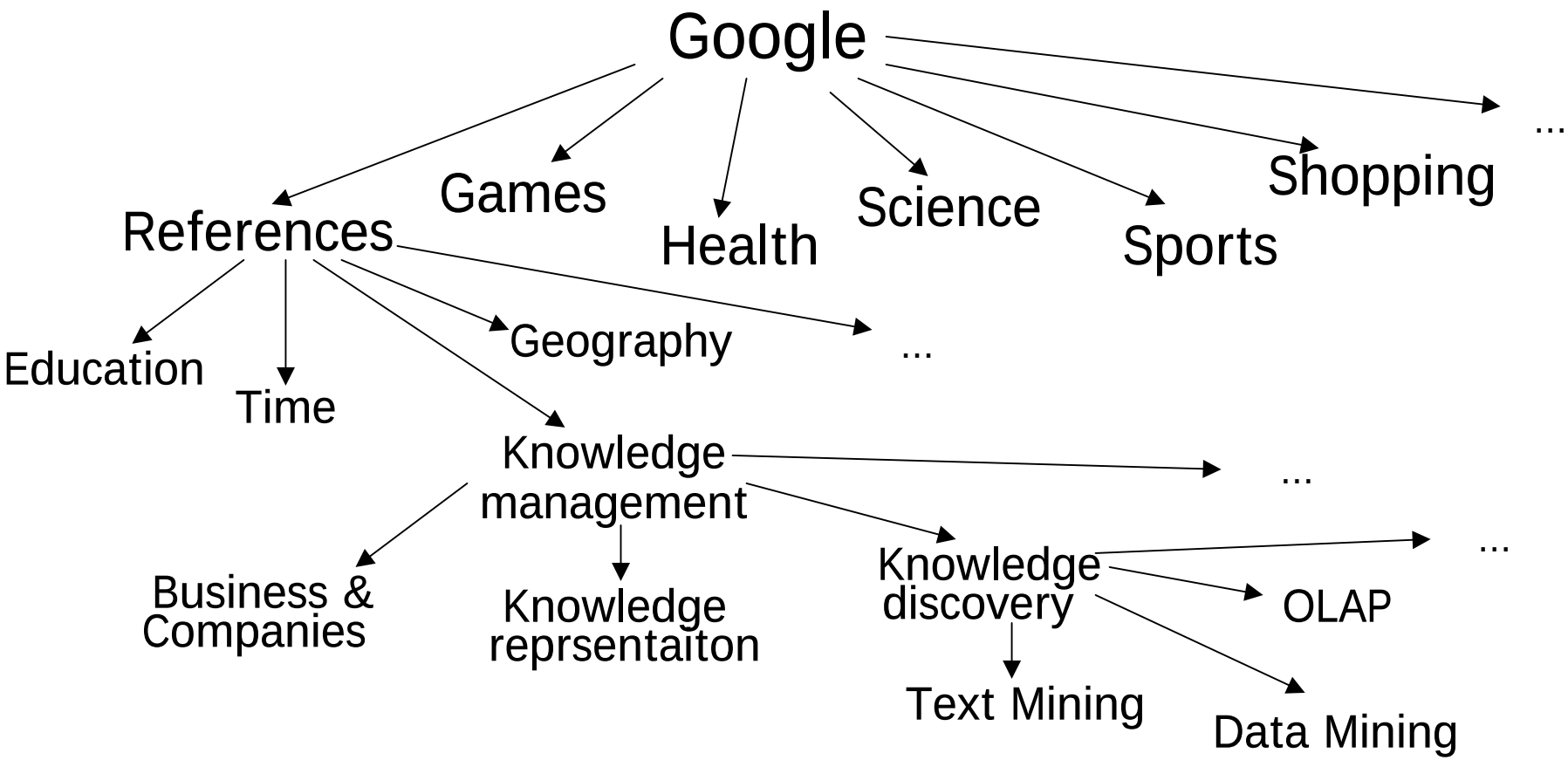
Extract relevant information by pictures, images, audio, video, spatial data... not only text.



Discovery in the Web graph

- Look at the Web as a graph where Web sites are subgraphs and many resources, as Web direstories, have a hierarchical structure.
- Use knowledge discovery algorithms for structured and graphical data to discover new interesting patterns.

Discovery in the Web graph: Google directories



Dealing with unexplored sizes

- Cover the gap between KD algorithms input size and Web space dimensions (10,000 against 10,000,000)
- KD algorithms should allow incremental refinement to manage the dynamic nature of the Web.
- KD algorithms should exploit hyper-links and access patterns.

References

- D. Malerba, F. Esposito, & M. Ceci (2002). Mining HTML pages to support document sharing in a cooperative system. In R. Unland, A. Chaudri, D. Chabane & W. Lindner (Eds.), *XML-Based Data Management and Multimedia Engineering - EDBT 2002 Workshops*, Lecture Notes in Computer Science, 2490, 420-434, Springer, Berlin, Germany.
- F. Sebastiani (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, Vol. 34, No. 1, 1-47.
- G. Paliouras (2003). Knowledge Discovery on the Web, Lectures during the Advanced Course in Information Technology and Society, Torino, Italy.
- D. Mladenic(1999). Text Learning and Related Intelligent Agents: A survey, *IEEE Intelligent System*, 44-54.
- R. Kosala, H. Blockeel (2000). Web Mining Research: A Survey*, ACM SIGKDD Vol.2 Issue 1,* 1-15, SIGKDD Explorations.