Term Projects

- A web-based tool for data mining
- Web Path Evaluator with Java
- Intelligent Solutions for Enterprise Web Server
- Data Mining**技術在網站上之應用** (e-MakeUp)
- Exploiting Data Mining in the Stock Market
- An Adaptive Multi-Attribute Multi-Measurement Method for Mining Classification Rules
- Mining Mobile Sequential Patterns in the Wireless E-Commerce Environment
- Mining the Most Interesting Association Rule
- Mining Relevant Patterns from Personal Mobility in a Mobile Comm. And Comput. Environment

1

Miscellaneous

- **Project presentation on 5/24**
- **Final Exam on 5/31**
- **Project due date 6/15**

2

Information Retrieval and Text Databases

- **Information retrieval:**
  - **IR: 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.**
- **Typical IR systems:**
  - **online library catalogs, online document management systems**
- **Information retrieval vs. database systems**
  - **Some DB problem not in IR, e.g., update, transaction management, complex objects.**
  - **Some IR problem not addressed well in DBMS, e.g., unstructured documents, approximate search using keywords and relevance.**

3

Challenges of Text Data Mining

- **Text database is ubiquitous:**
  - **library database, document database, e-mails, WWW, etc.**
- **The amount of text data increases very rapidly:**
  - **Text is easier to produce (news)**
  - **Networks allow easy access (Web is a huge text base).**
- **Problem: Information overhead**
  - **Analyst needs to find *right* information.**
- **Information retrieval is not enough**
  - **Too many documents that may contain useful information**
  - **Analyst may not even know what is needed without seeing documents (better retrieval not likely to help).**
  - **Problem may not be finding the right documents but patterns/trends across multiple documents.**

4

Text Database: Models and Retrieval Techniques

- **A simple model:**
  - **A document is represented by a string, which can be identified by a set of keywords.**
- **Major difficulties of the model:**
  - **Synonymy: A word T does not appear anywhere in the document, even though the document is closely related to T, e.g., data mining.**
  - **Polysemy: The same word may mean different things in different contexts, e.g., mining.**
- **Basic measures for content-based text retrieval**
  - ***Precision*: how many of the documents retrieved are in fact correct?**
  - ***Recall*: how many documents that should have been retrieved were in fact retrieved?**

5

Queries in Text Databases

- **Keyword-based information retrieval:**
  - **Popular not only in text database but also in multimedia data (e.g., video/audio clips, etc.)**
- **Queries may use *expressions* formed out of keywords:**
  - **E.g., car *and* repair shop, tea *or* coffee, DBMS *but not* Oracle**
  - **Queries and retrieval should consider synonyms: e.g., repair, maintenance**
- **Another types of query is: similarity-based retrieval (finding similar documents)**
  - **based on a set of common keywords**
- **Answer should be based on the degree of relevance**
  - **Relevance based on the nearness of the keywords, relative frequency of the keywords, etc.**

6

Basic Techniques in Text Retrieval Systems

- **Stop list:**
  - A text retrieval system often associates a stop list with a document set, which is a set of words that are deemed "irrelevant", e.g., *a, the, of, for, with*, etc., even though they may appear frequently.
  - Stop lists may vary when document set varies, e.g., "computer".
- **Word stem:**
  - Several words are small syntactic variants of each other since they share a common word stem, e.g., *drug*, *drugs*, *drugged*.
- **Frequency table:**
  - *Frequent_Table(I, j)*: # of occurrences of the word *t* in document *d*.
  - Usually, *ratio* instead of the absolute number of occurrences is used.
  - Measure the closeness of a document to a query (a set of keywords):
    - term distance

7

Text Retrieval Techniques

- **Inverted indices**
  - Two tables: DocTable (a set of Doc records) and TermTable
  - Doc record: <doc_id, postings_list (a list of terms or term pointers that occur in the document, sorted according to relevance)>.
  - Term record: <term, postings_list (a list of docs in which term appears)>
  - Answer query: Find all docs associated with one or a set of terms.
- **Signature files**
  - Associate a signature *Sd* with each document *d*.
  - A signature is a representation of an ordered list of terms that describe the document.
  - Order is obtained by frequency analysis, stemming and using stop lists.

8

Text Retrieval Applications:

- **Typical document retrieval queries:**
  - **Similarity of two documents**
  - **Finding the top p matches for a query Q.**
- **Some commercial systems:**
  - **Informix' datablade architecture provides a wide range of text database products that may be used in conjunction with their Universal Server's datablade architecture.**
    - **E.g, ArborText Document Objects, Open Text Livelink Library.**
  - **Oracle: ConText**
    - **Supports full text retrieval using SQL,**
    - **Automatically extracts themes from text and creates summaries**
  - **IBM DB2: Text Extender**
    - **Supports full text retrieval of multilingual documents, including search on keywords, synonyms, and word/phrase variations.**

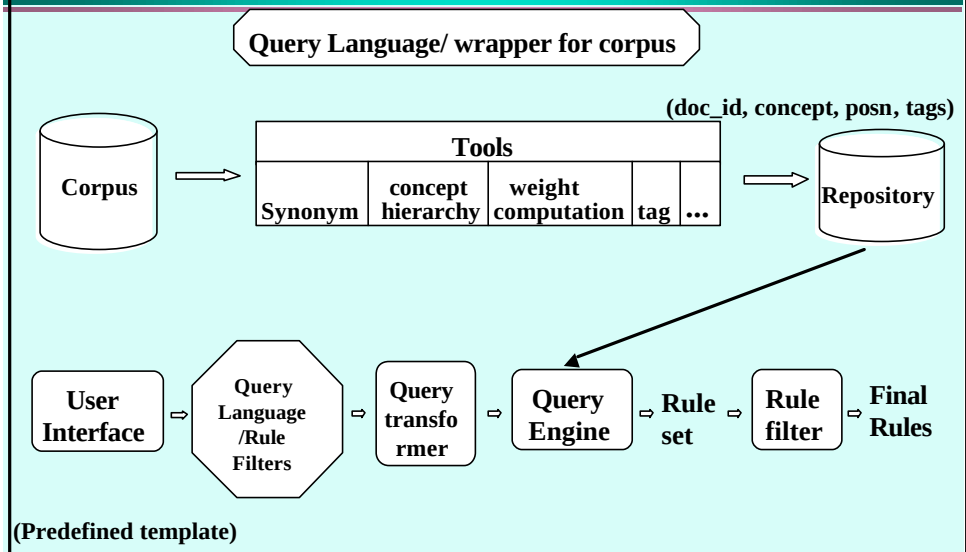
9

Types of Text Data Mining

- **Automatic document classification**
- **Link analysis:**
  - **Unusual correlation between entities.**
- **Sequence analysis:**
  - **What will predict a recurring event.**
- **Similarity detection:**
  - **Cluster documents by a common author**
  - **Cluster documents containing information from a common source.**
- **Anomaly detection:**
  - **Find information that violates usual patterns (unique source or miss information)**
- **Hypertext analysis:**
  - **Patterns in anchors/links**
  - **Anchor text correlations with linked objects.**

10

Text Mining System Architecture



11

Automatic Document Classification

- **Motivation:**
  - **Automatic classification of a large number of text documents (including Web pages, e-mails, or other types of files) based on a set of pre-classified training set.**
- **A classification problem:**
  - **Training set: Human experts generate a training data set**
  - **Classification: The computer system discovers the classification rules**
  - **Application: The discovered rules can be applied to classify new/unknown documents and put them into the right class.**

12

The EDF Project

- A research project at Electric de France
- Motivation: To classify a large number of projects defined each year (nearly 1,500) in EDF's research center (of 2,700 people), representing more than 2,000 pages of text
- The system was fully operational and an experimental study showed that the classifier is more reliable than the human experts who did the job before the classifier was built
- Reasons for the success
  - All available pieces of information are used
  - Traditional IR techniques are widely used in the system
  - The non-textual data have strong relationship with the correct class

13

The EDF Document Classification Methods

- A sample set of research project is classified by a human expert -- training set
- The title and text of each project is transformed into a set of significant words
- Discriminant analysis are performed on the two sets of keywords extracted
- A set of classification rules is generated for the classifier
- The classifier is augmented by rules learned from the non-textual data (such as department, customer, …)

14

The Singapore Web Document Classification Project

- Developed in National Univ. of Singapore (K. Wang, et al.'99)
- Major technology used:
  - Extract key words from text (Yahoo, ACM Web site)
  - Take the available classified documents as training set
  - Use multi-level association mining to find frequent sets
  - Ordering association rules based on the strength of rules --- Global classification instead of local classification
  - May handle synonyms, polysems, and distances between terms
  - An integration of association and classification
- High performance and high classification accuracy.

15

IBM's Intelligent Miner for Text

- Intelligent Miner for Text
  - a tool kit that provides a number of techniques upon which text mining based application can be easily built.
- Major tools included:
  - Extract key information from text
  - Organize document by subject
  - Find the predominant themes in a collection of documents
  - Search for relevant document using powerful and flexible queries --- Support multiple query model --boolean, fuzzy, free-text, …
- Mining functions
  - Extraction can be performed in individual features
  - Cluster search results
  - Refine queries (relevance feedback)

16

Major Components

- **Text analysis tools**
  - **Feature extraction -- annotating documents**
  - **Categorization -- organizing documents**
  - **Clustering -- document navigation**
- **Advanced search engines**
  - **Advanced text search engine -- TextMiner**
  - **Web-enabled search engine -- NetQuestion**
- **Web tools**
  - **Web Crawler**
  - **Web Crawler Toolkit**

17

Feature Extraction

- **To discover automatically the language(s) in which the document is written**
- **To recognize significant vocabulary items in text**
- **To recognize all names referring to a single entity**
- **To provide the location of all person names, places, and organization in the text**
- **To find multi-word terms that have a meaning of their own**
- **To find abbreviations introduced in a text and link them with their full names**

18

Categorization/Classification

- Users determine the taxonomy for organizing the documents into topics
- User create training sets to define categories
- Each document is analyzed and a rank value assigned as it relates to each category

19

Clustering

- To automatically group related documents based on their content
- Requires no training sets or predetermined taxonomies, generates a taxonomy at runtime
- Major steps:
  - Preprocessing
    - remove stop words, stem, feature extraction, lexical analysis, ...
  - Hierarchical clustering
    - computing similarities applying clustering algorithms, ..
  - Slicing
    - controls fan out, flattens the tree to configurable number of levels, ...

20

Clustering versus Categorization

- **Clustering: Documents in collection are processed and grouped into dynamically generated clusters**
- **Categorization: Documents in a collection are processed and grouped into predetermined groupings based on a taxonomy generated with a training set**

21

Interesting Applications

- **Text search by example**
- **Searching with categories**
- **Processing e-mail**

22

References

- C. Faloutsos. Access methods for text. ACM Comput. Surv., 17:49-74, 1985.
- R. Feldman and I. Dagan. Knowledge discovery in textual databases (KDT). Proc. 1st Int. Conf. Knowledge Discovery and Data Mining, Montreal, Canada, Aug. 1995.
- W. Frakes and R. Baeza-Yates. Information Retrieval: Data Structures and Algorithms. Printice Hall, 1992.
- V. Gaede and O. Gunther. Multdimensional access methods. ACM Comput. Surv., 30:170-231, 1998.
- L. Gravano, H. Garcia-Molina, and A. Tomasic. The effectiveness of gioss for the text database discovery problem. In SIGMOD'94.
- K. S. Jones and P. Willett (eds.). Readings in Information Retrieval, 3rd ed., Morgan Kaufmann, 1997.
- G. Salton. Automatic Text Processing. Addison-Wesley, 1989.
- G. Salton, J. Allen, C. Buckley, and A. Singhal. Automatic analysis, theme generation, and summarization of machine-readable texts. Science, 264:1421-1426, 1994.
- C. T. Yu and W. Meng. Principles of database query processing for advanced applications. Morgan Kaufmann, 1997.
- O. R. Zaiane, M. Xin, and J. Han. Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs, in ADL'98.
- C. Zaniolo, S. Ceri, C. Faloutsos, R. T. Snodgrass, C. S. Subrahmanian, and R. Zicari. Advanced database systems. Morgan Kaufmann, 1997.

23