Unit 8

Q.1 Explain text mining and discus in brief the information retrieval methods.

Q.2 Write a short note on text mining.

Text mining. Data mining can be applied to different kinds of texts, which represent unstructured data, in order to classify articles, books, documents, emails and web pages. Examples are web search engines or the automatic classification of press releases for storing purposes. Other text mining applications include the generation of filters for email messages and newsgroups.

Most previous studies of data mining have focused on structured data, such as relational, transactional, and data warehouse data. However, in reality, a substantial portion of the available information is stored in text databases (or document databases), which consist of large collections of documents from various sources, such as news articles, research papers, books, digital libraries, e-mail messages, and Web pages. Text databases are rapidly growing due to the increasing amount of information available in electronic form, such as electronic publications, various kinds of electronic documents, e-mail, and the World Wide Web (which can also be viewed as a huge, interconnected, dynamic text database). Nowadays most of the information in government, industry, business, and other institutions are stored electronically, in the form of text databases.
Data stored in most text databases are semistructured data in that they are neither completely unstructured nor completely structured. For example, a document may contain a few structured fields, such as title, authors, publication date, category, and so on, but also contain some largely unstructured text components, such as abstract and contents. There have been a great deal of studies on the modeling and implementation of semistructured data in recent database research. Moreover, information retrieval techniques, such as text indexing methods, have been developed to handle unstructured documents.

Traditional information retrieval techniques become inadequate for the increasingly vast amounts of text data. Typically, only a small fraction of the many available documents will be relevant to a given individual user. 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 need tools to compare different documents, rank the importance and relevance of the documents, or find patterns and trends across multiple documents. Thus, text mining has become an increasingly popular and essential theme in data mining.

Text Data Analysis and Information Retrieval

" Information retrieval (IR) is a field that has been developing in parallel with database systems for many years. Unlike the field of database systems, which has focused on query and transaction processing of structured data, information retrieval is concerned with the organization and retrieval of information from a large number of text-based documents.
Since information retrieval and database systems each handle different kinds of data, some database system problems are usually not present in information retrieval systems, such as concurrency control, recovery, transaction management, and update. Also, some common information retrieval problems are usually not encountered in traditional database systems, such as unstructured documents, approximate search based on keywords, and the notion of relevance.
Due to the abundance of text information, information retrieval has found many applications. There exist many information retrieval systems, such as on-line library catalog systems, on-line document management systems, and the more recently developed Web search engines.

Basic Measures for Text Retrieval: Precision and Recall
"Suppose that a text retrieval system has just retrieved a number of documents for me based on my input in the form of a query. How can we assess how accurate or correct the system was?" Let the set of documents relevant to a query be denoted as {Relevant}, and the set of documents retrieved be denoted as {Retrieved}. The set of documents that are both relevant and retrieved is denoted as {Relevant} ∩ {Retrieved}, as shown in the Venn diagram of Figure 10.6. There are two basic measures for assessing the quality of 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

$$\text{Precision} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Retrieved}\}|}$$

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

$$\text{Recall} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Relevant}\}|}$$

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:

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

The harmonic mean discourages a system that sacrifices one measure for another too drastically.

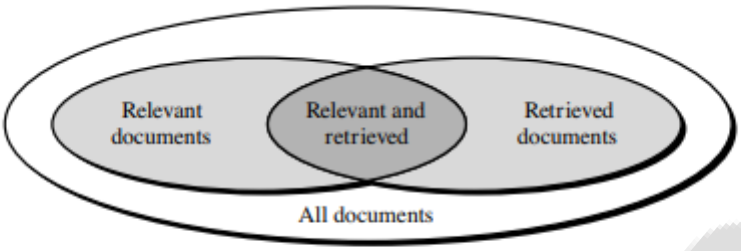

Figure 10.6 Relationship between the set of relevant documents and the set of retrieved documents.

Precision, recall, and F-score are the basic measures of a retrieved set of documents. These three measures are not directly useful for comparing two ranked lists of documents because they are not sensitive to the internal ranking of the documents in a retrieved set. In order to measure the quality of a ranked list of documents, it is common to compute an average of precisions at all the ranks where a new relevant document is returned. It is also common to plot a graph of precisions at many different levels of recall; a higher curve represents a better-quality information retrieval system. For more details about these measures, readers may consult an information retrieval textbook, such as [BYRN99].

Text Retrieval Methods

" Broadly speaking, retrieval methods fall into two categories: They generally either view the retrieval problem as a document selection problem or as a document ranking problem.

In document selection methods, the query is regarded as specifying constraints for selecting relevant documents. A typical method of this category is the Boolean retrieval model, in which a document is represented by a set of keywords and a user provides a Boolean expression of keywords, such as "car and repair shops," "tea or coffee," or "database systems but not Oracle." The retrieval system would take such a Boolean query and return documents that satisfy the Boolean expression. Because of the difficulty in prescribing a user's information need exactly with a Boolean query, the Boolean retrieval method generally only works well when the user knows a lot about the document collection and can formulate a good query in this way.

Document ranking methods use the query to rank all documents in the order of relevance. For ordinary users and exploratory queries, these methods are more appropriate than document selection methods. Most modern information retrieval systems present a ranked list of documents in response to a user's keyword query. There are many different ranking methods based on a large spectrum of mathematical foundations, including algebra, logic, probability, and statistics. The common intuition behind all of these methods is that we may match the keywords in a query with those in the documents and score each document based on how well it matches the query. The goal is to approximate the degree of relevance of a document with a score computed based on information such as the frequency of words in the document and the whole collection. Notice that it is inherently difficult to provide a precise measure of the degree of relevance between a set of keywords. For example, it is difficult to quantify the distance between data mining and data analysis. Comprehensive empirical evaluation is thus essential for validating any retrieval method.

Q.3 Write Notes on Web mining.

Q.4 Compare web content mining and web usage mining.

The web is a critical channel for the communication and promotion of a company's image. Moreover, e-commerce sites are important sales channels. Hence, it is natural to use web mining methods in order to analyze data on the activities carried out by the visitors to a website. Web mining methods are mostly used for three main purposes, as shown in Figure 13.14: content mining, structure mining and usage mining.

Web mining. Web mining applications, which will be briefly considered in section 13.1.9, are intended for the analysis of so-called clickstreams – the sequences of pages visited and the choices made by a web surfer. They may prove useful for the analysis of e-commerce sites, in offering flexible and customized pages to surfers, in caching the most popular pages or in evaluating the effectiveness of an e-learning training course.

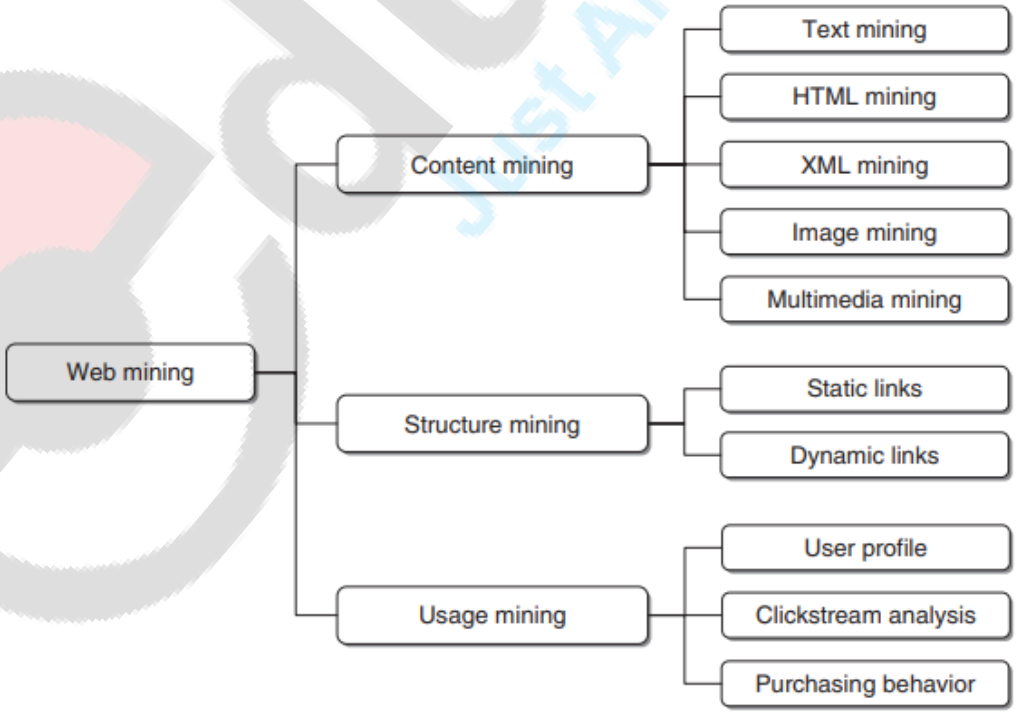

Figure 13.14 Taxonomy of web mining analyses

Content mining. Content mining involves the analysis of the content of web pages to extract useful information. Search engines primarily perform content mining activities to provide the links deemed interesting in relation to keywords supplied by users. Content mining methods can be traced back to data mining problems for the analysis of texts, both in free format or HTML and XML formats, images and multimedia content. Each of these problems is in turn dealt with using the learning models described in previous chapters. For example, text mining analyses are usually handled as multicategory classification problems, where the target variable is the subject category to which the text refers, while explanatory variables correspond to the meaningful words contained in the text.
Once it has been converted into a classification problem, text mining can be approached using the methods described in Chapter 10. Text mining techniques are also useful for analyzing the emails received by a support center. Notice that the input data for content mining analyses are easily retrievable, at least in principle, since they consist of all the pages that can be visited on the Internet.

Structure mining. The aim of this type of analysis is to explore and understand the topological structure of the web. Using the links presented in the various pages, it is possible to create graphs where the nodes correspond to the web pages and the oriented arcs are associated with links to other pages. Results and algorithms from graph theory are used to characterize the structure of the web, that is, to identify areas with a higher density of connections, areas disconnected from others and maximal cliques, which are groups of pages with reciprocal links. In this way, it is possible to pinpoint the most popular sites, or to measure the distance between two sites, expressed in terms of the lowest number of arcs along the paths that connect them in the links graph. Besides analyses aimed at exploring the global structure of the web, it is also possible to carry out local investigations to study how a single website is articulated. In some investigations, the local structure of websites is associated with the time spent by the users on each page, to verify if the organization of the site suffers from inconsistencies that jeopardize its effectiveness. For example, a page whose purpose is to direct navigation on the site should be viewed by each user only briefly. Should this not be the case, the page has a problem due to a possible ambiguity in the articulation of the links offered.

Usage mining. Analyses aimed at usage mining are certainly the most relevant from a relational marketing standpoint, since they explore the paths followed by navigators and their behaviors during a visit to a company website. Methods for the extraction of association rules are useful in obtaining correlations between the different pages visited during a session. In some instances, it is possible to identify

a visitor and recognize her during subsequent sessions. This happens if an identification key is required to access a web page, or if a cookie-enabling mechanism is used to keep track of the sequence of visits. Sequential association rules or time series models can be used to analyze the data on the use of a site according to a temporal dynamic. Usage mining analysis is mostly concerned with clickstreams – the sequences of pages visited during a given session. For e-commerce sites, information on the purchase behavior of a visitor is also available.

Q.5 Write a short note on Dimensionality reduction for text.

Dimensionality Reduction for Text

With the similarity metrics introduced in Section 10.4.1, we can construct similaritybased indices on text documents. Text-based queries can then be represented as vectors, which can be used to search for their nearest neighbors in a document collection. However, for any nontrivial document database, the number of terms T and the number of documents D are usually quite large. Such high dimensionality leads to the problem of inefficient computation, since the resulting frequency table will have size T × D. Furthermore, the high dimensionality also leads to very sparse vectors and increases the difficulty in detecting and exploiting the relationships among terms (e.g., synonymy). To overcome these problems, dimensionality reduction techniques such as latent semantic indexing, probabilistic latent semantic analysis, and locality preserving indexing can be used. We now briefly introduce these methods. To explain the basic idea beneath latent semantic indexing and locality preserving indexing, we need to use some matrix and vector notations. In the following part, we use x1,...,xtn ∈ R m to represent the n documents with m features (words). They can be represented as a term-document matrix X = [x1,x2,...,xn].

Latent Semantic Indexing Latent semantic indexing (LSI) is one of the most popular algorithms for document dimensionality reduction. It is fundamentally based on SVD (singular value
decomposition). Suppose the rank of the term-document X is r, then LSI decomposes X using SVD as follows:

$X = U\Sigma V^T$

where $\Sigma = diag(\sigma 1,...,\sigma r)$ and $\sigma 1 \geq \sigma 2 \geq \cdots \geq \sigma r$ are the singular values of X, U = [a1,...,ar] and ai is called the left singular vector, and V = [v1,...,vr], and vi is called

the right singular vector. LSI uses the first k vectors in U as the transformation matrix to embed the original documents into a k-dimensional subspace. It can be easily checked that the column vectors of U are the eigenvectors of XXT . The basic idea of LSI is to extract the most representative features, and at the same time the reconstruction error can be minimized. Let a be the transformation vector.
The objective function of LSI can be stated as follows:

$a_{opt} = \arg\min(a) \; || X - aa^T X ||^2 = \arg\max(a) \; a^T XX^T a,$

with the constraint, $a^T a = 1$.

Since XX^T is symmetric, the basis functions of LSI are orthogonal.

Locality Preserving Indexing
Different from LSI, which aims to extract the most representative features, Locality Preserving Indexing (LPI) aims to extract the most discriminative features. The basic idea of LPI is to preserve the locality information (i.e., if two documents are near each other in the original document space, LPI tries to keep these two documents close together in the reduced dimensionality space). Since the neighboring documents (data points in highdimensional space) probably relate to the same topic, LPI is able to map the documents related to the same semantics as close to each other as possible.
LSI aims to find the best subspace approximation to the original document space in the sense of minimizing the global reconstruction error. In other words, LSI seeks to uncover the most representative features. LPI aims to discover the local geometrical structure of the document space. Since the neighboring documents (data points in highdimensional space) probably relate to the same topic, LPI can have more discriminating power than LSI. Theoretical analysis of LPI shows that LPI is an unsupervised approximation of the supervised Linear Discriminant Analysis (LDA). Therefore, for document clustering and document classification, we might expect LPI to have better performance than LSI. This was confirmed empirically.

Probabilistic Latent Semantic Indexing
The probabilistic latent semantic indexing (PLSI) method is similar to LSI, but achieves dimensionality reduction through a probabilistic mixture model. Specifically, we assume there are k latent common themes in the document collection, and each is characterized by a multinomial word distribution. A document is regarded as a sample of a mixture model with these theme models as components. We fit such a mixture model to all the documents, and the obtained k component multinomial models can be regarded as defining k new semantic

dimensions. The mixing weights of a document can be used as a new representation of the document in the low latent semantic dimensions.