CHAPTER-26
Mining Text Databases

26.Mining Text Databases

26.1 Introduction

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 forms, such as electronic publications, e-mail, CD-ROMs and the World Wide Web (which can also be viewed as a huge, interconnected, dynamic text database).

Data stored in most text databases are semi-structured 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, length, 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 semi-structured 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 or 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.

26.2 Text Data Analysis and Information Retrieval

Information retrieval (TR) 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. A typical information retrieval problem is to locate relevant documents based on user input, such as keywords or example documents. Typical information retrieval systems include online library catalog systems and online document management systems.

Since information retrieval and database systems each handle different kinds of data, there are some database system problems that are usually not present in information retrieval systems, such as concurrency control, recovery, transaction management, and update unstructured documents, approximate search based on keywords, and the notion of relevance.

26.3 Basle Measures for Text Retrieval

“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]. 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

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

Recall:

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

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

26.4 Keyword-Based and Similarity-Based Retrieval

Most information retrieval systems support keyword-based and/or similarity-based retrieval. In keyword- based information retrieval, a document is represented by a string, which can be identified by a set of keywords. A user provides a keyword or an expression formed out of a set of keywords, such as "car and repair shops", "tea or coffee", or "database systems but not Oracle". A good information retrieval system should consider synonyms when answering such queries. For example, given the keyword car, synonyms such as automobile and vehicle should be considered in the search as well. Keyword-based retrieval is a simple model that can encounter two major difficulties. The first is the synonymy problem: a keyword, such as software product, may not appear anywhere in the document, even though the document is closely related to software product. The second is the polysemy problem: the same keyword, such as mining, may mean different things in different contexts.

Similarity-based retrieval finds similar documents based on a set of common keywords. The output of such retrieval should be based on the *degree of relevance*, where relevance is measured based on the closeness of the keywords, the relative frequency of the keywords, and so on. Notice that in many cases, it is difficult to provide a precise measure of the degree of relevance between a set of keywords, such as the distance between data mining and data analysis.

A text retrieval system often associates a sop list with a set of documents. A stop list is a set of words that are deemed "irrelevant." For example, a, the, of, for, with, and so on are stop words even though they may appear frequently. Stop lists may vary when the document sets vary. For example, database systems

could be an important keyword in a newspaper. However, it may be considered as a stop word in a set of research papers presented in a database systems conference.

A group of different words may share the same word stem. A text retrieval system needs to identify groups of words where the words in a group are small syntactic variants of one another, and collect only the common **word stem** per group. For example, the group of words drug, drugged, and drugs, share a common word stem, drug, and can be viewed as different occurrences of the same word.

26.5 Other Text Retrieval Indexing Techniques

There are several other popularly adopted text retrieval indexing techniques, including inverted indices and signature files.

An inverted index is an index structure that maintains two hash indexed or B+-tree indexed tables: document_table and term_table, where

- *document_table* consists of a set of document records, each containing two fields: doc_id and posting_list, where posting_list is a list of terms (or pointers to terms) that occur in the document, sorted according to some relevance measure.
- *term_table* consists of a set of term records, each containing two fields: term_id here posting_list specifies a list of document identifiers in which the term appears.

With such organization, it is easy to answer queries like "Find all of the documents associated with a given set of terms ", or "Find all of the terms associated with a givers set of documents". For example, to find all of the documents associated with a set of terms, we can first find a list of document identifiers in term_table for each term, and then intersect them to obtain the set of relevant documents. Inverted indices are widely used in industry. They are easy to implement, but are not satisfactory at handling synonymy and polysemy. The postmg_lists could be rather long, making the storage requirement quite large.

A **signature file** is a file that stores a signature record for each document in the database. Each signature has a fixed size of b bits representing terms. A simple encoding scheme goes as follows. Each bit of a document signature is initialized to 0. A bit is set to -1 if the term it represents appears in the document, A Signature S1 matches another signature S2 if each bit that is set in signature S2 is also set in S1. Since there are usually more terms than available bits, there may be multiple terms mapped into the same bit. Such multiple-to-one mappings make the search expensive since a document that matches the signature of a query does not necessarily contain the set of keywords of the query. The document has to be retrieved, parsed, stemmed, and checked. Improvements can be made by first performing frequency analysis, stemming, and by filtering stop words, and then using a hashing technique and superimposed coding technique, to encode the list of terms into bit representation. Nevertheless, the problem of multiple-to-one mappings still exists, which is the major disadvantage of this approach.

26.6 Text Mining: Keyword-Based Association and Document Classification

"What about mining associations in text databases? Can we also generate document classification schemes?" This subsection addresses both of these questions.

Keyword-Based Association Analysis

"What is keyword-based association analysis?" Such analysis collects sets of keywords or terms that occur frequently together and then finds the association or correlation relationships among them.

Like most of the analyses in text databases, association analysis first preprocesses the text data by parsing, stemming, removing stop words, and so on, and then evokes association-mining algorithms. In a document database, each document can be viewed as a transaction, while a set of keywords in the document can be considered as a set of items in the transaction. That is, the database is in the format

[document_id, a_set_of_keywords].

The problem of keyword association mining in document databases is thereby mapped to item association mining in transaction databases, where many interesting methods have been developed.

Notice that a set of frequently occurring consecutive or closely located key-words may form a term or a phrase. The association mining process can help detect compound associations, that is, domain-dependent terms or phrases, such as [Stanford, University] or [U.S. President, Bill, Clinton], or non-compound associations, such as [dollars, shares, exchange, total, commission, stake, securities].

Mining based on these associations is referred to as “term level association mining” (as opposed to mining on individual words). Term recognition and term level association mining enjoy two advantages in text analysis: (1) terms and phrases are automatically tagged so that there is no need for human effort in tagging documents, and (2) the number of meaningless results is greatly reduces, as is the execution time of the mining algorithms.

With such term and phrase recognition, term level mining can be evoked to find associations among a set of detected terms and keywords. Some users may like to find associations between pairs of keywords or terms from a given set of keywords or phrases, whereas others may wish to find the maximal set of terms occurring together. Therefore, based on user mining requirements, standard association mining or max-pattern mining algorithms may be evoked.

Document Classification Analysis

Automated document classification is an important text-mining task since, with the existence of a tremendous number of on-line documents, it is tedious yet essential to be able to automatically organize such documents into classes so as to facilitate document retrieval and subsequent analysis.

A general procedure for performing document classification is as follows: First, a set of pre-classified documents is taken as the training as the training set. The training set is then analyzed in order to derive a classification scheme. Such a classification scheme often needs to be refined with a testing process. The so-derived classification scheme can be used for classification of other on-line documents.

This process appears similar to the classification of relational data. However, there is a fundamental difference. Relational data are well structured: each tuple is defined by a set of attribute-value pairs. For example, in the tuple [*sunny, warm, dry, not_windy, play_tennis*], the value “sunny” corresponds to the attribute weather_outlook, “warm” corresponds to the attribte temperature, and so on. The classification analysis decides which set of attribute-value pairs has the greatest discriminating power in determining whether or not a person is going to play tennis. On the other hand, document databases are

not structured according to attribute-value pairs. That is, a set of keywords associated with a set of documents is not organized into a fixed set of attributes or dimensions. Therefore, commonly used relational data-oriented classification methods, such as decision tree analysis, cannot be used to classify document databases.

An effective method for document classification is to explore association-based classification, which classifies documents based on a set of associated, frequently occurring text patterns. Such an association –based classification method proceeds as follows: First, keywords and terms can be extracted by information retrieval and simple association analysis techniques. Second, concept hierarchies of keywords and terms can be obtained using available term classed, such as WordNet, or relying on expert knowledge, or some keyword classification systems. Documents in the training set can also be classified into class hierarchies. A term association mining method can then be applied to discover sets of associated terms that can be used to maximally distinguish one class of documents from others. This derives a set of association rules associated with each document class. Such classification rules can be ordered based on their occurrence frequency and discriminative power, and used to classify new documents. Such a kind of association-based document classifier has been proven effective. For Web document classification of document classes. Web linkage analysis methods are discussed in the next section.

26.7 Review Questions

1 Explain about Text Data Analysis and Information Retrieval

2 Explain about Basle Measures for Text Retrieval

3 Explain about Keyword-Based and Similarity-Based Retrieval

4 Explain about Other Text Retrieval Indexing Techniques

5 Explain about Text Mining: Keyword-Based Association and Document Classification

26.8 References.

[1]. Data Mining Techniques, Arun k pujari 1st Edition

[2] .Data warehousung,Data Mining and OLAP, Alex Berson ,smith.j. Stephen

[3].Data Mining Concepts and Techniques ,Jiawei Han and MichelineKamber

[4]Data Mining Introductory and Advanced topics, Margaret H Dunham PEA

[5] The Data Warehouse lifecycle toolkit , Ralph Kimball Wiley student Edition