



Review of Types of Data Used for Data Mining

Vipan

PG Department of Computer Science
A.S. College, Khanna, Punjab, India

Robin Kumar

PG Department of Mathematics
A.S. College, Khanna, Punjab, India

Abstract - This Paper is dedicated to the study and analysis of different types of data that can be used to for data mining i.e. object, spatial, multimedia, text, and Web data. And also discusses some challenges in data mining. These kinds of data are commonly encountered in many social, economic, scientific, engineering, and governmental applications, and pose new challenges in data mining. Mining of different types of data has different types of challenges in this IT era. Data mining tools are developed and improved day by day as the size of the data is growing tremendously at faster pace and it becomes more difficult to extract valuable information from such large amount of data of deferent type that is available these days for data mining or extraction of knowledge.

Keywords: data mining, Knowledge extraction, spatial data, multimedia data, object data.

I. INTRODUCTION

Data mining refers to extracting or mining knowledge from large amounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining should have been more appropriately named “knowledge mining from data,” which is unfortunately somewhat long. “Knowledge mining,” a shorter term, may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. Thus, such a misnomer that carries both “data” and “mining” became a popular choice. Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery.

As we have large amount of data for knowledge extraction and have different types of data such as discussed in further sections.

II. COMPLEX DATA OBJECTS

Many advanced, data-intensive applications, such as scientific research and engineering design, need to store, access, and analyze complex but relatively structured data objects. These objects cannot be represented as simple and uniformly structured records (i.e., tuples) in data relations. Such application requirements have motivated the design and development of *object-relational* and *object-oriented* database systems. Both kinds of systems deal with the efficient storage and access of vast amounts of disk-based complex structured data objects. These systems organize a large set of complex data objects into *classes*, which are in turn organized into *class/subclass* hierarchies. Each object in a class is associated with (1) an *object-identifier*, (2) a *set of attributes* that may contain sophisticated data structures, set- or list-valued data, class composition hierarchies, multimedia data, and (3) a *set of methods* that specify the computational routines or rules associated with the object class. There has been extensive research in the field of database systems on how to efficiently index, store, access, and manipulate complex objects in object-relational and object-oriented database systems. Technologies handling these issues are discussed in many books on database systems, especially on object-oriented and object-relational database systems.

One step beyond the storage and access of massive-scaled, complex object data is the systematic analysis and mining of such data. This includes two major tasks: (1) construct multidimensional data warehouses for complex object data and perform online analytical processing (OLAP) in such data warehouses, and (2) develop effective and scalable methods for mining knowledge from object databases and/or data warehouses. The second task is largely covered by the mining of specific kinds of data (such as spatial, temporal, sequence, graph- or tree-structured, text, and multimedia data), since these data form the major new kinds of complex data objects.

III. SPATIAL DATA MINING

Spatial data mining refers to the extraction of knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial databases. Such mining demands an integration of data mining with spatial database technologies. It can be used for understanding spatial data, discovering spatial relationships and relationships between

spatial and non spatial data, constructing spatial knowledge bases, reorganizing spatial databases, and optimizing spatial queries. It is expected to have wide applications in geographic information systems, geo marketing, remote sensing, image database exploration, medical imaging, navigation, traffic control, environmental studies, and many other areas where spatial data are used. A crucial challenge to spatial data mining is the exploration of *efficient* spatial data mining techniques due to the huge amount of spatial data and the complexity of spatial data types and spatial access methods. A spatial database stores a large amount of space-related data, such as maps, pre processed remote sensing or medical imaging data, and VLSI chip layout data. Spatial databases have many features distinguishing them from relational databases. They carry topological and/or distance information, usually organized by sophisticated, multidimensional spatial indexing structures that are accessed by spatial data access methods and often require spatial reasoning, geometric computation, and spatial knowledge representation techniques.

As with relational data, we can integrate spatial data to construct a data warehouse that facilitates spatial data mining. A spatial data warehouse is a *subject-oriented, integrated, time-variant, and non volatile* collection of both spatial and non spatial data.

There are several challenging issues regarding the construction and utilization of spatial data warehouses. The first challenge is the integration of spatial data from heterogeneous sources and systems. Spatial data are usually stored in different industry firms and government agencies using various data formats. Data formats are not only structure-specific (e.g., raster- vs. vector-based spatial data, object-oriented vs. relational models, different spatial storage and indexing structures), but also vendor-specific (e.g., ESRI, MapInfo, Intergraph). There has been a great deal of work on the integration and exchange of heterogeneous spatial data, which has paved the way for spatial data integration and spatial data warehouse construction.

The second challenge is the realization of fast and flexible on-line analytical processing in spatial data warehouses. The star schema model is a good choice for modeling spatial data warehouses because it provides a concise and organized warehouse structure and facilitates OLAP operations. However, in a spatial warehouse, both dimensions and measures may contain spatial components.

There are three types of *dimensions* in a spatial data cube: Nonspatial dimension, Spatial-to-nonspatial dimension, Spatial-to-spatial dimension.

IV. MULTIMEDIA DATA MINING

A multimedia database system stores and manages a large collection of *multimedia data*, such as audio, video, image, graphics, speech, text, document, and hypertext data, which contain text, text markups, and linkages. Multimedia database systems are increasingly common owing to the popular use of audio video equipment, digital cameras, CD-ROMs, and the Internet. Typical multimedia database systems include NASA's EOS (Earth Observation System), various kinds of image and audio-video databases, and Internet databases.

Our study of multimedia data mining focuses on image data mining. Here we introduce multimedia data mining, including similarity search in multimedia data, multidimensional analysis, classification and prediction analysis, and mining associations in multimedia data.

A. Similarity Search in Multimedia Data

When we searching for similarities in multimedia data, the question arises that, can we search on either the data description or the data content? That is correct. For similarity searching in multimedia data, we consider two main families of multimedia indexing and retrieval systems: (1) *description-based retrieval systems*, which build indices and perform object retrieval based on image descriptions, such as keywords, captions, size, and time of creation; and (2) *content-based retrieval systems*, which support retrieval based on the image content, such as color histogram, texture, pattern, image topology, and the shape of objects and their layouts and locations within the image. Description-based retrieval is labor-intensive if performed manually. If automated, the results are typically of poor quality. For example, the assignment of keywords to images can be a tricky and arbitrary task. Recent development of Web-based image clustering and classification methods has improved the quality of description-based Web image retrieval, because image surrounded text information as well as Web linkage information can be used to extract proper description and group images describing a similar theme together.

Content-based retrieval uses visual features to index images and promotes object retrieval based on feature similarity, which is highly desirable in many applications.

Several approaches have been proposed and studied for similarity-based retrieval in image databases, based on image signature:

- Color histogram-based signature
- Multifeature composed signature
- Wavelet-based signature
- Wavelet-based signature with region-based granularity

B. Multidimensional Analysis of Multimedia Data

To facilitate the multidimensional analysis of large multimedia databases, multimedia data cubes can be designed and constructed in a manner similar to that for traditional data cubes from relational data. A multimedia data cube can contain additional dimensions and measures for multimedia information, such as color, texture, and shape. Here we

examine a multimedia data mining system prototype called MultiMediaMiner, which extends the DBMiner system by handling multimedia data. The example database tested in the MultiMediaMiner system is constructed as follows. Each image contains two descriptors: a *feature descriptor* and a *layout descriptor*. The original image is not stored directly in the database; only its descriptors are stored. The description information encompasses fields like image file name, image URL, image type (e.g., gif, tiff, jpeg, mpeg, bmp, avi), a list of all known Web pages referring to the image (i.e., parent URLs), a list of keywords, and a thumbnail used by the user interface for image and video browsing. The feature descriptor is a set of vectors for each visual characteristic. The main vectors are a color vector containing the color histogram quantized to 512 colors (8X8X8 for *RGB*), an MFC (Most Frequent Color) vector, and an MFO (Most Frequent Orientation) vector. The MFC and MFO contain five color centroids and five edge orientation centroids for the five most frequent colors and five most frequent orientations, respectively. A multimedia data cube can have many dimensions. The following are some examples: the size of the image or video in bytes; the width and height of the frames (or pictures), constituting two dimensions; the date on which the image or video was created (or last modified); the format type of the image or video; the frame sequence duration in seconds; the image or video Internet domain; the Internet domain of pages referencing the image or video (parent URL); the keywords; a color dimension; an edge-orientation dimension; and so on.

The multimedia data cube seems to be an interesting model for multidimensional analysis of multimedia data. However, we should note that it is difficult to implement a data cube efficiently given a large number of dimensions. This curse of dimensionality is especially serious in the case of multimedia data cubes.

C. Classification and Prediction Analysis of Multimedia Data

Classification and predictive modelling have been used for mining multimedia data, especially in scientific research, such as astronomy, seismology, and geo scientific research. In general, all of the classification methods can be used in image analysis and pattern recognition. Moreover, in-depth statistical pattern analysis methods are popular for distinguishing subtle features and building high-quality models. *Data preprocessing* is important when mining image data and can include data cleaning, data transformation, and feature extraction. Aside from standard methods used in pattern recognition, such as edge detection and Hough transformations, techniques can be explored, such as the decomposition of images to eigenvectors or the adoption of probabilistic models to deal with uncertainty. Since the image data are often in huge volumes and may require substantial processing power, parallel and distributed processing are useful. Image data mining classification and clustering are closely linked to image analysis and scientific data mining, and thus many image analysis techniques and scientific data analysis methods can be applied to image data mining.

The popular use of the World Wide Web has made the Web a rich and gigantic repository of multimedia data. The Web not only collects a tremendous number of photos, pictures, albums, and video images in the form of on-line multimedia libraries, but also has numerous photos, pictures, animations, and other multimedia forms on almost every Web page. Such pictures and photos, surrounded by text descriptions, located at the different blocks of Web pages, or embedded inside news or text articles, may serve rather different purposes, such as forming an inseparable component of the content, serving as an advertisement, or suggesting an alternative topic. Furthermore, these Web pages are linked with other Web pages in a complicated way. Such text, image location, and Web linkage information, if used properly, may help understand the contents of the text or assist classification and clustering of images on the Web. Data mining by making good use of relative locations and linkages among images, text, blocks within a page, and page links on the Web becomes an important direction in Web data analysis.

D. Mining Associations in Multimedia Data

Association rules involving multimedia objects can be mined in image and video databases. At least three categories can be observed: Associations between image content and non image content features: A rule like “*If at least 50% of the upper part of the picture is blue, then it is likely to represent sky*” belongs to this category since it links the image content to the keyword *sky*. Associations among image contents that are not related to spatial relationships: A rule like “*If a picture contains two blue squares, then it is likely to contain one red circle as well*” belongs to this category since the associations are all regarding image contents. Associations among image contents related to spatial relationships: A rule like “*If a red triangle is between two yellow squares, then it is likely a big oval-shaped object is underneath*” belongs to this category since it associates objects in the image with spatial relationships. To mine associations among multimedia objects, we can treat each image as a transaction and find frequently occurring patterns among different images.

There are some differences between mining association rules in multimedia databases versus in transaction databases. First, an image may contain multiple objects, each with many features such as color, shape, texture, keyword, and spatial location, so there could be many possible associations. In many cases, a feature may be considered as the same in two images at a certain level of resolution, but different at a finer resolution level. Therefore, it is essential to promote a progressive resolution refinement approach. That is, we can first mine frequently occurring patterns at a relatively rough resolution level, and then focus only on those that have passed the minimum support threshold when mining at a finer resolution level. This is because the patterns that are not frequent at a rough level cannot be frequent at finer resolution levels. Such a multi resolution mining strategy substantially reduces the overall data mining cost without loss of the quality and completeness of data mining results.

This leads to an efficient methodology for mining frequent item sets and associations in large multimedia databases. Second, because a picture containing multiple recurrent objects is an important feature in image analysis, recurrence of the same objects should not be ignored in association analysis. For example, a picture containing two golden circles is

treated quite differently from that containing only one. This is quite different from that in a transaction database, where the fact that a person buys one gallon of milk or two may often be treated the same as “buys milk.” Therefore, the definition of multimedia association and its measurements, such as support and confidence, should be adjusted accordingly. Third, there often exist important spatial relationships among multimedia objects, such as *above*, *beneath*, *between*, *nearby*, *left-of*, and so on. These features are very useful for exploring object associations and correlations. Spatial relationships together with other content-based multimedia features, such as color, shape, texture, and keywords, may form interesting associations. Thus, spatial data mining methods and properties of topological spatial relationships become important for multimedia mining.

E. Audio and Video Data Mining

Besides still images, an incommensurable amount of audiovisual information is becoming available in digital form, in digital archives, on the World Wide Web, in broadcast data streams, and in personal and professional databases. This amount is rapidly growing. There are great demands for effective content-based retrieval and data mining methods for audio and video data. Typical examples include searching for and multimedia editing of particular video clips in a TV studio, detecting suspicious persons or scenes in surveillance videos, searching for particular events in a personal multimedia repository such as MyLifeBits, discovering patterns and outliers in weather radar recordings, and finding particular melody or tune in your MP3 audio album. To facilitate the recording, search, and analysis of audio and video information from multimedia data, industry and standardization committees have made great strides toward developing a set of standards for multimedia information description and compression. For example, MPEG-*k* (developed by MPEG: *Moving Picture Experts Group*) and JPEG are typical video compression schemes. The most recently released MPEG-7, formally named “*Multimedia Content Description Interface*,” is a standard for describing the multimedia content data. It supports some degree of interpretation of the information meaning, which can be passed onto, or accessed by, a device or a computer. MPEG-7 is not aimed at any one application in particular; rather, the elements that MPEG-7 standardizes support as broad a range of applications as possible. The audiovisual data description in MPEG-7 includes still pictures, video, graphics, audio, speech, three-dimensional models, and information about how these data elements are combined in the multimedia presentation.

Video data mining is still in its infancy. There are still a lot of research issues to be solved before it becomes general practice. Similarity-based pre processing, compression, indexing and retrieval, information extraction, redundancy removal, frequent pattern discovery, classification, clustering, and trend and outlier detection are important data mining tasks in this domain.

V. TEXT MINING

Most 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 modelling 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.

VI. MINING THE WORLD WIDE WEB

The World Wide Web serves as a huge, widely distributed, global information service center for news, advertisements, consumer information, financial management, education, government, e-commerce, and many other information services. The Web also contains a rich and dynamic collection of hyperlink information and Web page access and usage information, providing rich sources for data mining. However, based on the following observations, the Web also poses great challenges for effective resource and knowledge discovery.

The Web seems to be too huge for effective data warehousing and data mining. The size of the Web is in the order of hundreds of terabytes and is still growing rapidly. Many organizations and societies place most of their public-accessible information on the Web. It is barely possible to set up a data warehouse to replicate, store, or integrate all of the data on the Web.

The complexity of Web pages is far greater than that of any traditional text document collection. Web pages lack a unifying structure. They contain far more authoring style and content variations than any set of books or other traditional text-based documents. The Web is considered a huge digital library; however, the tremendous number of documents in this library are not arranged according to any particular sorted order.

There is no index by category, nor by title, author, cover page, table of contents, and so on. It can be very challenging to search for the information you desire in such a library! *The Web is a highly dynamic information source.* Not only does the Web grow rapidly, but its information is also constantly updated. News, stock markets, weather, sports, shopping, company advertisements, and numerous other Web pages are updated regularly on the Web. Linkage information and access records are also updated frequently. *The Web serves a broad diversity of user communities.* The Internet currently connects more than 100 million workstations, and its user community is still rapidly expanding. Users may have very different backgrounds, interests, and usage purposes. Most users may not have good knowledge of the structure of the information network and may not be aware of the heavy cost of a particular search. They can easily get lost by groping in the “darkness” of the network, or become bored by taking many access “hops” and waiting impatiently for a piece of information. *Only a small portion of the information on the Web is truly relevant or useful.* It is said that 99% of the Web information is useless to 99% of Web users. Although this may not seem obvious, it is true that a particular person is generally interested in only a tiny portion of the Web, while the rest of the Web contains information that is uninteresting to the user and may swamp desired search results. How can the portion of the Web that is truly relevant to your interest be determined? How can we find high quality Web pages on a specified topic?

These challenges have promoted research into efficient and effective discovery and use of resources on the Internet. There are many index-based Web search engines. These search the Web, index Web pages, and build and store huge keyword-based indices that help locate sets of Web pages containing certain keywords. With such search engines, an experienced user may be able to quickly locate documents by providing a set of tightly constrained keywords and phrases. However, a simple keyword-based search engine suffers from several deficiencies. First, a topic of any breadth can easily contain hundreds of thousands of documents. This can lead to a huge number of document entries returned by a search engine, many of which are only marginally relevant to the topic or may contain materials of poor quality. Second, many documents that are highly relevant to a topic may not contain keywords defining them. This is referred to as the *polysemy* problem, discussed in the previous section on text mining. For example, the keyword *Java* may refer to the Java programming language, or an island in Indonesia, or brewed coffee. As another example, a search based on the keyword *search engine* may not find even the most popular Web search engines like Google, Yahoo!, AltaVista, or America Online if these services do not claim to be search engines on their Web pages. This indicates that a simple keyword based Web search engine is not sufficient for Web resource discovery.

If a keyword-based Web search engine is not sufficient for Web resource discovery, how can we even think of doing Web mining? Compared with keyword-based Web search, Web mining is a more challenging task that searches for Web structures, ranks the importance of Web contents, discovers the regularity and dynamics of Web contents, and mines Web access patterns. However, Web mining can be used to substantially enhance the power of A Web search engine since Web mining may identify authoritative Web pages, classify Web documents, and resolve many ambiguities and subtleties raised in keyword-based Web search. In general, Web mining tasks can be classified into three categories: *Web content mining*, *Web structure mining*, and *Web usage mining*. Alternatively, Web structures can be treated as a part of Web contents so that Web mining can instead be simply classified into *Web content mining* and *Web usage mining*.

VII. CONCLUSION

In this paper, briefly discuss the basic concept of the data mining, challenging and application areas for data mining. And mainly focused on the different types of data included in data mining with various forthcoming challenges in extraction of valuable information and knowledge from complex data such as object oriented data, multimedia data which is available tremendously in this era and increases its size at a faster pace. Data mining is more than running some complex queries on the data you stored in your database. Identifying the format of the information that you need is based upon the technique, type of data and the analysis that you want to do. To Analyze, manage and make a decision of such type of huge amount of complex data we need newer techniques for data mining which will transforming in many fields. Data mining should be applicable to any kind of information repository. The challenges listed by different types of data very significantly.

REFERENCES

- [1] Osmer R. Zalane, CMPUT 690 principles of knowledge discovery in databases”, *Introduction to Data mining*”.
- [2] M.S. Chen, J. Han and P.S. Yu. “*Data mining : An overview from a database perspective*”. IEEE transactions on Knowledge and data engineering 8:866.
- [3] “*Introduction to Data Mining and Knowledge Discovery*” , Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.
- [4] Larose, D. T., “*Discovering Knowledge in Data: An Introduction to Data Mining*”, ISBN 0-471-66657-2, John Wiley & Sons, Inc, 2005.
- [5] Dunham, M. H., Sridhar S., “*Data Mining: Introductory and Advanced Topics*”, Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006.
- [6] Tan Pang-Ning, Steinbach, M., Vipin Kumar. “*Introduction to Data Mining*”, Pearson Education, New Delhi.

- [7] Campos, M. M., Stengard, P. J., Boriana, L. M., “*Data-Centric Automated Data Mining*”, WebSite: www.oracle.com/technology/products/bi/odm/pdf/automated_data_mining_paper_1205.pdf
- [8] Xingquan Zhu, Ian Davidson, “*Knowledge Discovery and Data Mining: Challenges and Realities*”, ISBN 978-1-59904-252, Hershey, New York, 2007.
- [9] Zhao, Kaidi and Liu, Bing, Tirpark, Thomas M. And Weimin, Xiao, “*A Visual Data Mining Framework for Convenient Identification of Useful Knowledge*”, ICDM '05 Proceedings of the Fifth IEEE International Conference on Data Mining, vol.-1, no.-1, pp.- 530-537, Dec 2005.
- [10] Jiawei Han, Champaign Micheline Kamber, “*Data Mining: Concepts and Techniques*”, University of Illinois at Urbana-Champaign.