Journal of Data Mining and Knowledge Discovery
ISSN: 2229–6662 & ISSN: 2229–6670, Volume 3, Issue 1, 2012, pp.-40-44.
Available online at http://www. bioinfo. in/contents. php?id=42

SPATIAL DATA CLASSIFICATION AND DATA MINING

RATHI J.B. * AND PATIL A.D.

Department of Computer Science & Engineering, Jawaharlal Darda Institute of Engineering & Technology, Maharashtra- 445001, India.
*Corresponding Author: Email- jyoti.rathi08@gmail.com

Received: February 21, 2012; Accepted: March 15, 2012

Abstract- Text categorization is the problem of classifying text documents into a set of predefined classes. Text categorization algorithms usually represent documents as bags of words and consequently have to deal with huge numbers of features. The project work will implement text classification and mining. Semantic Analysis will be used for feature extraction, eliminating the text representation errors caused by synonyms and polysemesand reducing the dimension of text vector. At the same time, a new decision-making approach based on concentration, will be implemented for improving the classification of texts in overlapping regions .
Keywords- Data Mining, Spatial Data.

Citation: Rathi J. B. and Patil A. D. (2012) Spatial Data Classification and Data Mining. Journal of Data Mining and Knowledge Discovery, ISSN: 2229–6662 & ISSN: 2229–6670, Volume 3, Issue 1, pp. -40-44.

Introduction

Information plays vital role in every field. To generate information it requires massive collection of data. Data can be from simple numerical measurements and text documents, to more complex information such as spatial data, multimedia dataand hypertext documents. To take complete advantage of data; the data retrieval is simply not enough, it requires a tool for automatic summarization of data, extraction of the essence of information storedand the discovery of patterns in raw data . With the enormous amount of data stored in files, databases and other repositories, it is increasingly important, to develop powerful tool for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. 'Data mining' is the only tool to accomplish the above mentioned needs.

Data mining is the extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors. Data mining tools answer the questions that traditionally were too time consuming to resolve. They prepare databases for finding hidden patterns, finding predictive information.

Data mining is popularly known as Knowledge Discovery in Databases (KDD), it refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases.

A. The Data Mining Tasks

The data mining tasks are of different types, data mining perform on data store with the aim to generate knowledge which can be used to take decisions in the business or any other systems. Depending on the use of data mining result the data mining tasks are classified into following types:

1. Exploratory Data Analysis: It is simply exploring the data without any clear ideas of what we are looking for. These techniques are interactive and visual.
2. Descriptive Modeling: It describe all the data, it includes models for overall probability distribution of the data, partitioning of the n-dimensional space into groups and models describing the relationships between the variables.
3. Predictive Modeling: This model permits the value of one variable to be predicted from the known values of other variables.
4. Discovering Patterns and Rules: It concern with pattern detec-

tion, the aim is spotting fraudulent behavior by detecting regions of the space defining the different types of transactions where the data points significantly different from the rest.
5. Retrieval by Content: It is finding pattern similar to the pattern of interest in the data set. This task is most commonly used for text and image data sets.

B. Dimensions of Data Mining Context

There are four dimensions of the data mining context:
1. The application domain: It is the specific area of subject/ knowledge in which the data mining project takes place.
2. The data mining problem type: It describes the specific classes of objectives that the data mining project deals with.
3. The technical aspect: It covers specific issues in data mining that describe different technical challenges that usually occur during data mining.
4. The tool and technique required: It specifies which data mining tools and/or techniques are applied during the data mining project.

C. Data Mining Life Cycle

The life cycle of a data mining project consists of six phases. The sequence of the phases is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase. The main phases are: business understanding, data understanding, data preparation, modeling, evaluation and deployment.
1. Business Understanding: This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.
2. Data Understanding: The data understanding phase starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.
3. Data Preparation: The data preparation phase covers all activities to construct the final dataset from the initial raw data. The Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools.
4. Modeling: In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. There are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often necessary.
5. Evaluation: In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be certain it properly achieves the business objectives. A key objective is to determine whether any important business issue has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.
6. Deployment: The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

D. The Knowledge Discovery Process

Data mining is one of the tasks in the process of knowledge discovery from the database. The data stored in the database is used to discover the patterns of data, which then interpreted by applying the domain knowledge. Following figure shows the process of Knowledge Discovery from Database. The steps in the KDD process contain mainly data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and the knowledge representation.
1. Data cleaning: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.
2. Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.
3. Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
4. Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
5. Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
6. Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures.
7. Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

E. Data Mining Methods

The data mining methods are broadly categories as: On-Line Analytical Processing (OLAP), Classification, Clustering, Association Rule Mining, Temporal Data Mining, Time Series Analysis, Spatial Mining, Web Mining etc. The data source can be data warehouse, database, flat file or text file. The selection of data mining algorithm is mainly depends on the type of data used for mining and the expected outcome of the mining process.

The Intelligent Discovery Assistants (IDA) helps users applying valid knowledge discovery processes. The IDA can provide users with three facilities:
1. A systematic enumeration of valid knowledge discovery processes.
2. Effective rankings of valid processes by different criteria, which help to choose between the options.
3. An infrastructure for sharing knowledge, which leads to network externalities.

Literature Review

A. Spatial Data Mining

When the data has relations with spatial data, the term becomes spatial data mining. It will follow along the same functions in data mining; with the end objective is to find patterns in geography, meteorology etc.

Spatial data mining is the process of extracting implicit knowledge,

spatial relations, or other patterns that are not explicitly stored in spatial databases. For instance, spatial data mining can be used in spatial (statistical) analysis to very large datasets such as finding cancer clusters to locate hazardous environment or to find new spatial and interesting patterns such as find locations that are unusual etc. Spatial patterns of interest here may include characterization of locations of a feature (e. g. crime) and its association with other spatial features (e. g. population density, distance to transportation network, etc).

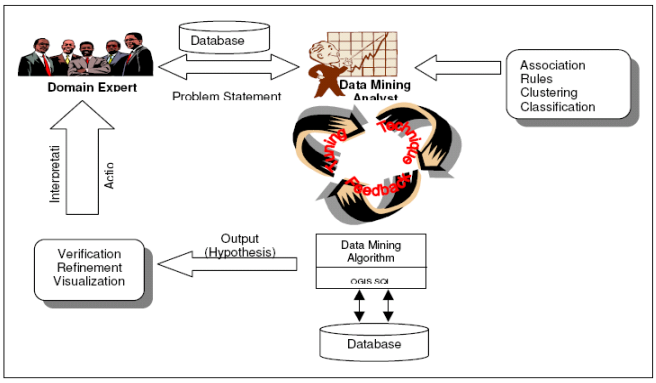


Fig. 1- A data mining process

The total number of spatial objects in the database sometimes can come up to a very large number. So the discovery process for spatial data becomes more complex than conventional data. Parallel with that extracting interesting pattern from traditional numeric and categorized data due to the complexity of spatial data types, spatial relationship and spatial autocorrelation. This applies to both the efficiency of algorithms and the complexity of possible patterns that can be found in a spatial database.

B. Spatial Analysis

Currently, moving object information can only be used for monitoring purpose and not be used for analysis. This drawback gives problem to the company or organization in order to identify type of pattern to respond. It is more critical if we need to known about emergency pattern where quick respond is the most important thing. For example in order to know information about moving object data (e. g. ambulance) in a database, the answer to the question of moving objects position (e. g. "How far is the ambulance with registered plate number MAM9777 from Street A?") needs continuously to be updated. When the database unable to handle or manage a dynamic attribute, cause the query that refer to future values of dynamic attributes impossible to answer. It is because to answer to a query, it is not only depends on the database content, but also on the time at which the query was requested.

Problem Definition

The domain/problem specific applications are designed by considering two parameters: Data to be mined and the purpose of mining. The use of data mining method depending on requirement changes form problem to problem. The context factor is also an important parameter along with data to be mined and method to be used for mining.

A. General Analysis of The Problem

Researchers in data mining and knowledge discovery are creating new, more automated methods for discovering knowledge. A thorough analysis of the problem required for deploying a data mining solution. The models need to be built, used and integrated with different applications. Employing common data mining standards simplifies the integration, updatingand maintenance of the applications and systems containing the models.

B. Problem Analysis of Data Mining Tools

The existing data mining tools are having limitations. During this following problems are identified in the existing data mining tools:

1. Difficult to use - Existing data mining tools try to cover all different data mining applications, thus it becomes very difficult to configure and run.
2. Needs Expert to run the tool - No domain or problem specific logic is tied with the tool, therefore needs expert to run the tool and analyze the result.
3. Difficult to add new functionality - Because of the size and complexity of each tool, it is very difficult to add any new feature.
4. Difficult to interface - There is no way to integrate algorithms developed by some other companies.
5. Short Lifetime - Changing the exiting tool to incorporate new feature is difficult and require lot of changes. With time the tool become obsolete, as new tools take the market.
6. Limited Number of algorithms – Existing tool only provides limited number of algorithm and sometime use of multiple algorithms is very limited.
7. Need lot of resources: Existing tools are not optimized for any specific application; therefore they need lot of resources, such as runtime memory, hard disk etc.

C. Analysis of Performance

In this, study of different data mining systems it is observed that the data mining engines are generic and domain/problem specific. The performance of different data mining engines is compared and it is observed that the domain/problem specific data mining applications has shown good performance over the generic engines.

There are several domain/problem specific applications have shown performance above 90%and the generic applications have limitations in its functioning therefore it is concluded that domain/problem specific applications are high performance solutions. It is therefore taken the problem of designing a domain/problem specific data mining engine which addresses some of the problems identified.

Implementation

The problem related literature survey is presented in the following sections.

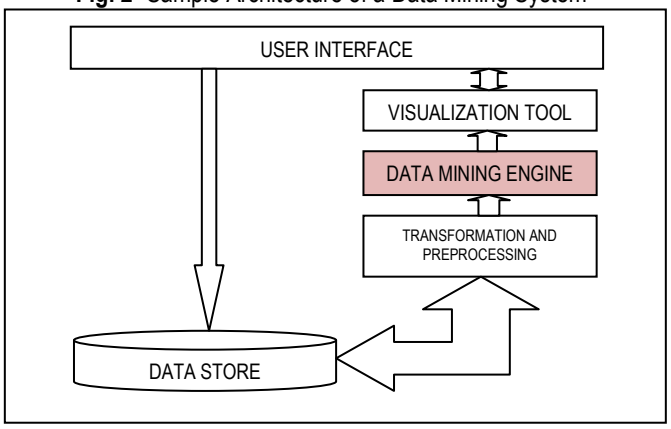
A. Sample Data Mining System Architecture

During study it is found that, generalized data mining engines have limitations. The users are required to apply his/her domain knowledge and skills to use the system.

Above figure of sample architecture depicts the main components of data mining system. User interface contains forms that enable

users to add data into data store and present the output.

Fig. 2- Sample Architecture of a Data Mining System



add data into data store and present the output. Optionally the user interface may be used to select data mining methods and algorithms. The Transformation and Preprocessing component transform data in the format suitable and accessible to Data Mining Engine and also provides mechanism for data preprocessing. Preprocessing includes mainly removal of noise from data and irrelevant data from the collection. The Data mining component includes the methods and algorithms for data mining. The visualization tool convert and prepares the output of data mining into suitable visual format and send it to user interface to present it to the users.

B. Generalized Data Mining Engines

Several attempts have been made to design the data mining tool that can autonomously select data, select the mining algorithm, retrieve the unknown patterns and interprets the outcome. The attempt is for minimizing the domain expertise required for mining of data. The designs proposed by different researchers found partially successful. Domain expertise is thus has no exception. The domain specific and problem oriented data mining applications have shown the maximum success rate.

Advantages & Application

Advantages

This is basically based on the concept of text categorization. Text categorization is the problem of classifying text documents into a set of predefined classes. Text categorization algorithms usually represent documents as bags of words and consequently have to deal with huge numbers of features. This work implements text classification and mining. Semantic Analysis is used for feature extraction, eliminating the text representation errors caused by synonyms and polysemesand reducing the dimension of text vector. At the same time, a new decision-making approach based on concentration, is implemented for improving the classification of texts in overlapping regions.

This application will be very user friendly. By using this application we can save a lot of time in reading the information which is not of our use.

Application

In general purpose the concept of data mining is widely used over the web applications. In this case, we are mining the data from a particular text file using the concept of data mining. So this application can be widely used over the internet to mine the data over the several web pages and make a profile as per the need of user.

Conclusion

Most of the domain specific data mining applications have shown accuracy above 90%. The generic data mining applications are having the limitations. The domain experts play important role at different stages of data mining. The decisions at different stages are influenced by the factors like domain and data details, aim of the data miningand the context parameters. The domain specific applications are aimed to extract specific knowledge. The results yields from the domain specific applications are more accurate and useful. Therefore it is concluded that the domain specific applications are more proper for data mining.

From work it is observed that, it is very difficult to design and develop a data mining system, which works dynamically for any domain. Designing a domain specific system is also a challenging task. Designing the Data mining engine that works for specific problem domain is thus the central idea of this study.

Limitations & Future Scope

Future Scope

The generic and problem specific data mining applications were studied during this. It is observed that the problem and domain specific data mining engines gave good and promising results. The system architecture presented in this study is implemented for Text Mining to generate Entity Profile. The initial knowledge generation from the selected corpus of text is done by using association rule mining algorithm. The results obtained have shown that the architecture presented in this for the problem and domain specific application, when implemented for the problem of generating players profile from the text given good results. The future scope for the study is as given below:

In future work can be extended to improve the results of the algorithms which have been used for further analysis and understanding of documents.

This work may also be extended to Mining of various data type and generate variety of knowledge.

The interfaces shall be design by using guidelines and techniques mentioned in this study. The interfaces shall be more interactive and guide and provide opportunity to the users to take decisions and select most proper option.

The system design needs further refinement by using expert system design principles. The knowledge base needs to be design to incorporate other form of the knowledge.

Limitation

Based on the architecture, scope of this research are the process of integration spatiotemporal data types in DBMS and producing the pattern by mining the spatio-temporal data in DBMS environment. Researcher has to find solution for integration spatiotemporal data types in DBMS environment as it is the basic steps in finding the pattern. Currently in the actual spatio-temporal data type it does not exist and this is the reason why we need to do the

integration before we could find solution in finding the pattern. After completing the integration process, we could proceed with the second process in finding the pattern by analyzing moving point object data. GPS log files are the source files for spatio-temporal data types. However to be more reliable, we decide to find collocation pattern only with the moving point object data meaning the new data type is a point object with the time. If this research becomes successful, further research should look into more objects in spatial types and more than one pattern or the generic pattern.

References

[1] Yahaya Abd. Rahim and Shahrin Sahib, *@utem. edu. my*.
[2] Dunham M. H. and Sridhar S. (2006) *Pearson Education*, *New Delhi, 1st Edition.*
[3] Pang-Ning T., Steinbach M. and Vipin Kumar (2009) *Pearson Education*, *New Delhi, 3rd Edition*.
[4] *Introduction to Data Mining and Knowledge Discovery* (1999) Third Edition.
[5] Berson A. and Smith S. J. (2007) *Data Warehousing, Data Mining & OLAP*.
[6] Berson A., Smith S. and Thearling K. (2000) *Building Data Mining Application for CRM*.
[7] Jensen Christian S. URL-*http://www. cs. aau. dk/~csj/Thesis/ pdf/chapter1. pdf*.
[8] Hand D., Heikki, M. and Smyth P. (2001) *Principles of Data Mining*, *New Delhi, 1st Edition*.
[9] Baazaoui-Zghal H., Faiz S. and Ben G. H. (2005) *World Academy of Science, Engineering and Technology,* 5.