Volume : 2 | Issue : 3 | March 2013 ISSN - 2250-1991

Research Paper Computer Science

Analyzing, Developing and Implementing Data Mining Techniques on Databases, Web Contents and Textual Data

*Ronak S. Raiyani **Dr. Bankim Radadiya
***Dr. Satish Thumar

* Research Scholar, “Shree Darshan”, 26/13 Bhojrajpara, Gondal – 360311

** Research Guide, Office of Directorate of Information Technology, ASPEE Agribusiness Management Institute, Navsari Agricultural University, Navsari - 396450

* Subject Expert, 8420 W 131st PL, #911, Overland Park KS – 66213, USA

ABSTRACT

The profusion of resources on the electronic media and storage capacity increase gave rise to considerable interest in the research community. Traditional information retrieval techniques have been applied to the document collection on the Internet, and panoply of search engines and tools have been proposed and implemented. However, the effectiveness of these tools is not satisfactory. None of them is capable of discovering knowledge from the Internet.

In this work we propose a data warehouse model on top of the existing infrastructure for implementing data-mining model with model for web and text mining for any kind of web pages or text data.

Large collections of data is gathered for a myriad of applications. Data mining from such a data corpus lead to interesting discoveries. For the web mining part, We have used many different types of web sites. And for text mining, We have taken the data from SEC EDGAR.

Keywords :

Objective

The objective is to explore how the data mining interprets database information and “how this information is used to organize actions”. This has created a special interest in making comparison of different algorithms of data mining with effort to develop experimental paradigms that allow testing the mining algorithms.

The data mining strategies do not follow the same track but show different pictures in different situations. The variability may make the implementation complex and success rate is not upto the anticipated level. To smoothen the track, the efforts are made to model the track of data mining that is expected to cover many different situations.

Major Characteristics of Research

- **Design Strategies**

o Naturalistic Inquiry
o Emergent Design Flexibility

- **Data-Collection and Fieldwork Strategies**

o Qualitative Data
o Personal experience and engagement
o Empathic neutrality and engagement
o Dynamic systems

- **Analysis Strategies**

o Uniquie case orientation
o Inductive analysis and creative synthesis
o Holistic perspective
o Context Sensitivity

Overview of KDD

The term Knowledge Discovery in Databases abbreviated as KDD refers to the broad process of extracts knowledge from huge data, and emphasizes the "high-level" application of particular data mining methods. The unifying goal of the KDD process is to extract knowledge from data in the context of large databases.

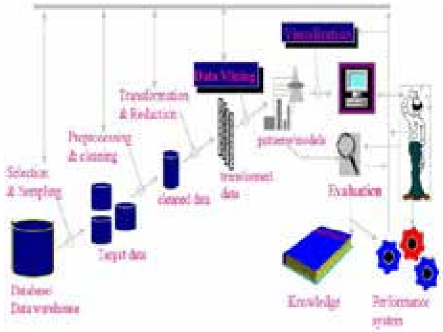

Fig 1. Process of finding and interpreting patterns from Data.

Understanding Data Warehouse

An Enterprise Data Warehouse (EDW) is an architectural component that is subject oriented, integrated, volatile and time variant. Data warehouse exist to enhance managements ability to make decision. There are some other important characteristics of data in the data warehouse. Data in the warehouse is granular. This means that data is carried in the data warehouse at the lowest level of granularity. For data warehousing efficient hardware and software architecture should be in place. Proper system architecture viz three tier/ two tier should be implemented

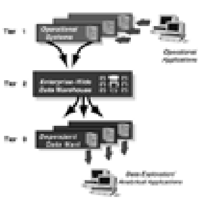

Fig 2. Three Tier Architecture

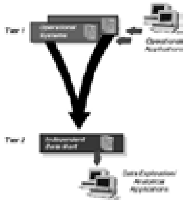

Fig 3. Two Tier Architecture

Data Mining

Large amount of data generated by organizations worldwide is mostly unorganized. If data is organized one can generate/ extract meaningful and useful information to convert unorganized data into organized data. Normally the concept of DBMS is implemented though a database in management systems is embedded with a query language popularly known as SQL server. The use of SQL particularly in unorganized large databank is not always adequate to meat the end user requirements.

Data mining is the technique of abstracting meaningful information form large and unorganized databanks. It involves the process of performing automated abstraction and generating predictive information from large databanks. The abstraction of meaningful large databanks can also be known as knowledge discovery. The data mining process uses of varieties of analysis tools to determine the relationship between data and the databank and to use the same to make valid prediction. Data mining techniques are a result of integration of various techniques forms multiple disciplines such as statistic, machine learning, pattern recognition, neural networks, image processing and so on. Fig 4. Below shows general phase of Data mining process.

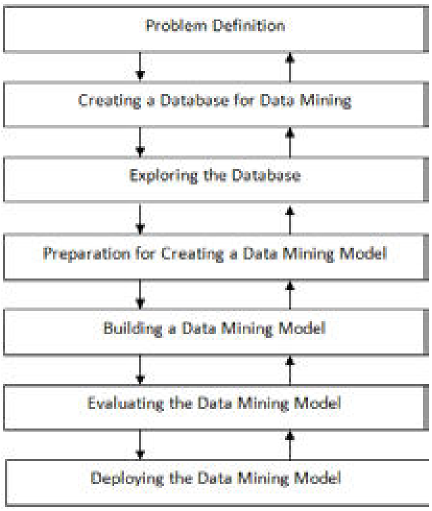
Problem Definition
Creating a Database for Data Mining
Exploring the Database
Preparation for Creating a Data Mining Model
Building a Data Mining Model
Evaluating the Data Mining Model
Deploying the Data Mining Model

Fig 4. General phase of Data Mining Process.

Data Mining Techniques

Data mining techniques provide a way to use data mining tasks in order to predict solution sets for a problem and a level of confidence about the predicted solution interns of consistency of prediction and interns of frequency of correct predictions. Data mining techniques include:

- Statistics
- Machine Learning
- Decision Tree
- Hidden markov Model
- Artificial Neural Network
- Genetic Algorithms
- Meta Learning
- Association Rules
- Clustering Techniques

Text Mining

Apart of Databases, Text format is a common and natural form of storing information. Text data are inherently unstructured and fuzzy. Thus text mining is more complicated process than general data mining. Domain of text Mining overlaps with several other fields, such as computational linguistics, and machine learning.

The study of the computerized applications and techniques, such as automatic machine translation and text analysis in processing and analyzing a language is commonly known as computational linguistics. In other words, computational linguistics is simply a branch of linguistics studies that applies computers for the research of linguistics. Text mining is widely used in computational linguistics.

A text-mining tool can be used to explore and analyze the content of textual documents and to visually display the extracted information with a graphical interface commonly known as a conceptual map or concept map, or sometimes document map. A concept map provides a close view of the material, representing the main concepts within the text and how they are related to each other. A content map also displays the conceptual structure of the extracted information. You can search for number of occurrences of concepts and their interrelations in the text documents. Text mining tools offers both quantifying and displaying the conceptual structure of a document set.

Text mining tools, such as SAS Text miner and Leximancer are used for various text-mining tasks. For example, you are creating an index of keywords used in a book. You need to know number of occurrences of a particular term along with the page numbers. Using a text-mining tool you can count number of occurrences of keywords. Text mining can be used for:

- Text analyzing: Enable you to exploring large text data and analyzing it using charts and graphs.
- Creating indexes of documents: Enables you to prepare document index
- Detecting plagiarism: Enable you to detect plagiarism. You can explore the content of a book or a large text and check for plagiarism.
- Statistical analysis of documents: Provides you various statistics regarding texts. For example, you can use a text-mining tool to count number of occurrences of a keyword in a single or across several documents.
- Document mapping: Can read a large text for you and create a map of the document collection. This does not necessarily mean that a text-mining tool understands the details of a text as well as you do. Example of a document mapping tasks is arranging parts of texts under appropriate heading.
- Processing customer letters: Enables you to summarize and categorize the customer queries and complain letters according to types of query or complain.
- Building archives of electronic data: Indicates creating a digital library, or news achieves.

Web Mining

Web mining is a specialized application of data mining. In simple words, Web mining is a technique to process information available on Web and search for useful data. Web mining enables you to discover Web pages, text documents, multimedia files, images and other types of resources from web. Pattern extraction is a Web mining process to monitor the original or uploaded web pages, extract information from them and generate matches of a specific pattern with necessary information specified by a user. The pattern extraction process enables you to efficiently surf and access data available on the Web:

Web mining is widely used in several fields. The various fields where Web mining is applied are:

- E-commerce
- Information filtering
- Fraud detection
- Plagiarism detection
- Education and research

The general techniques and algorithms of data mining are also applicable in Web mining Web mining tasks can be decomposed into four subtasks:

- Resource searching: Indicates the task of retrieving documents from the Web.

- Information selection: Denotes automatic extraction of information from the Web documents. Several Web mining tools such as Web Miner are available to perform this task.
- Generalization of patterns: Denotes automatic discovery of patterns across multiple web sites.
- Analysis of Web documents: Denotes validation and analysis of the extracted patterns.

REFERENCES

Rhonda Delmater, Monte Hancock . A Manager's Guide to customer-centric business intelligence: Digital Press. | Pieter Adriaans, Dolf Zantinge. Data Mining | Sam Anahory, Dennis Murray. A Practicle guide for Business DSS. | Internet Sites. http://www.kenorrinst.com/dwpaper | http://www.thearling.com/text/dmwhite/dmwhite.htm | http://intelligent-web.org/wsm/overview/ | http://www.sas.com/technologies/analytics/datamining/