International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 1

ISSN: 2454-4248
397 – 400

Brief Introduction of Data Mining and Data Warehousing

Rasika R. Wadegoankar, Prof. Reena Thakur
Computer Science & Engineering Guru Nanak Institute of Technology India

Abstract:- Over the past two decades there has been a huge increase in the amount of data being stored in databases as well as the number of database applications in business and the scientific domain. This explosion in the amount of electronically stored data was accelerated by the success of the relational model for storing data and the development and maturing of data retrieval and manipulation technologies. While technology for storing the data developed fast to keep up with the demand, little stress was paid to developing software for analysing the data until recently when companies realized that hidden within these masses of data was a resource that was being ignored. The huge amount of stored data contains knowledge about a number of aspects of their business waiting to be harnessed and used for more effective business decision support. Database Management Systems used to manage these data sets at present only allow the user to access information explicitly present in the databases i.e. the data. Contained implicitly within this data is knowledge about a number of aspects of their business waiting to be harnessed and used for more effective business decision support. This extraction of knowledge from large data sets is called Data Mining or Knowledge Discovery in Databases and is defined as the non-trivial extraction of implicit, previously unknown and potentially useful information from data. The obvious benefit of Data Mining has resulted in a lot of resources being directed towards its development.

*****

I. INTRODUCTION

Fundamentally, data mining is about processing data and identifying patterns and trends in that information so that you can decide or judge. Data mining principles have been around for many years, but, with the advent of *big data*, it is even more prevalent.

Big data caused an explosion in the use of more extensive data mining techniques, partially because the size of the information is much larger and because the information tends to be more varied and extensive in its very nature and content. With large data sets, it is no longer enough to get relatively simple and straightforward statistics out of the system. With 30 or 40 million records of detailed customer information, knowing that two million of them live in one location is not enough. You want to know whether those two million are a particular age group and their average earnings so that you can target your customer needs better.

These business-driven needs changed simple data retrieval and statistics into more complex data mining. The business problem drives an examination of the data that helps to build a model to describe the information that ultimately leads to the creation of the resulting report.

The process of data analysis, discovery, and model-building is often iterative as you target and identify the different information that you can extract. You must also understand how to relate, map, associate, and cluster it with other data to produce the result. Identifying the source data and formats, and then mapping that information to our given result can change after you discover different elements and aspects of the data.

II. MATERIALS AND METHODS

Data mining tools

Data mining is not all about the tools or database software that you are using. You can perform data mining with comparatively modest database systems and simple tools, including creating and writing your own, or using off the shelf software packages. Complex data mining benefits from the past experience and algorithms defined with existing software and packages, with certain tools gaining a greater affinity or reputation with different techniques.

It is recent that the very large data sets and the cluster and large-scale data processing are able to allow data mining to collate and report on groups and correlations of data that are more complicated. Now an entirely new range of tools and systems available, including combined data storage and processing systems.

You can mine data with a various different data sets, including, traditional SQL databases, raw text data, key/value stores, and document databases. Clustered databases, such as Hadoop, Cassandra, CouchDB, and Couchbase Server, store and provide access to data in such a way that it does not match the traditional table structure.

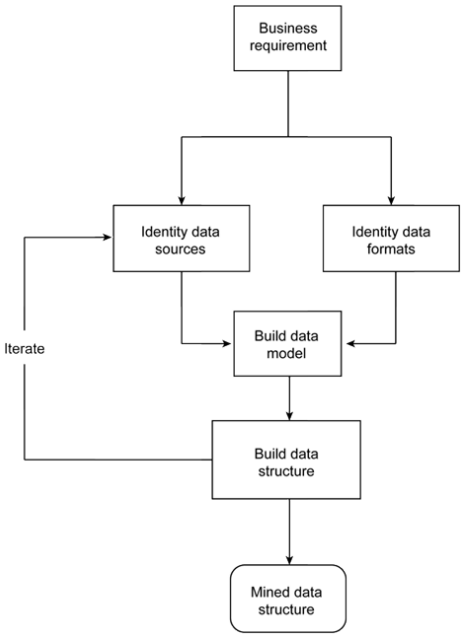
In particular, the more flexible storage format of the document database causes a different focus and complexity in terms of processing the information. SQL databases impost strict structures and rigidity into the schema, which makes querying them and analyzing the data straightforward from the perspective that the format and structure of the information is known.

Document databases that have a standard such as JSON enforcing structure, or files that have some machine-readable structure are also easier to process, although they

 ISSN: 2454-4248
397 – 400

might add complexities because of the differing and variable structure. For example, with Hadoop's entirely raw data processing it can be complex to identify and extract the content before you start to process and correlate.

Figure:

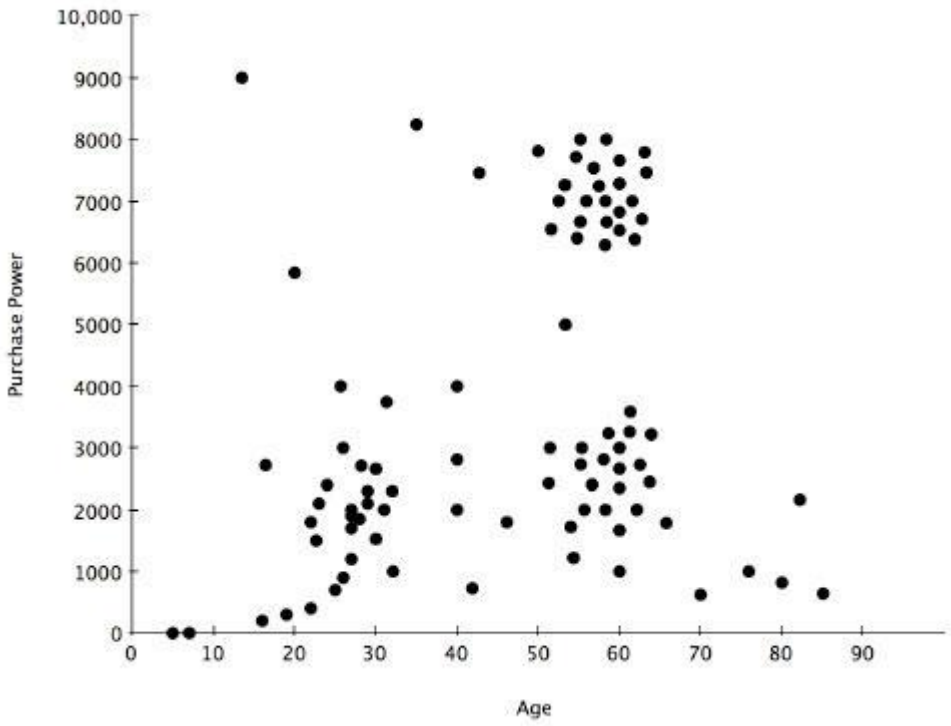


Outline of the process

The process of data analysis, discovery, and model-building is often iterative as you target and identify the different information that you can extract. You must also understand how to relate, map, associate, and cluster it with other data to produce the result. Identifying the source data and formats, and then mapping that information to our given result can change after you discover different elements and aspects of the data.

Clustering

By examining one or more attributes or classes, you can group individual pieces of data together to form a structure opinion. At a simple level, clustering is using one or more attributes as your basis for identifying a cluster of correlating results. Clustering is useful to identify different information because it correlates with other examples so you can see where the similarities and ranges agree.Clustering can work both ways. You can assume that there is a cluster at a certain point and then use our identification criteria to see if you are correct. The graph below showgood example. In this example, a sample of sales data compares the age of the customer to the size of the sale. It is not unreasonable to expect that people in their twenties (before marriage and kids), fifties, and sixties (when the children have left home), have more disposable income.



Prediction

Prediction is a wide topic and runs from predicting the failure of components or machinery, to identifying fraud and even the prediction of company profits. Used in combination with the other data mining techniques, prediction involves analyzing trends, classification, pattern matching, and relation. By analyzing past events or instances, you can make a prediction about an event.Using the credit card

 ISSN: 2454-4248
397 – 400

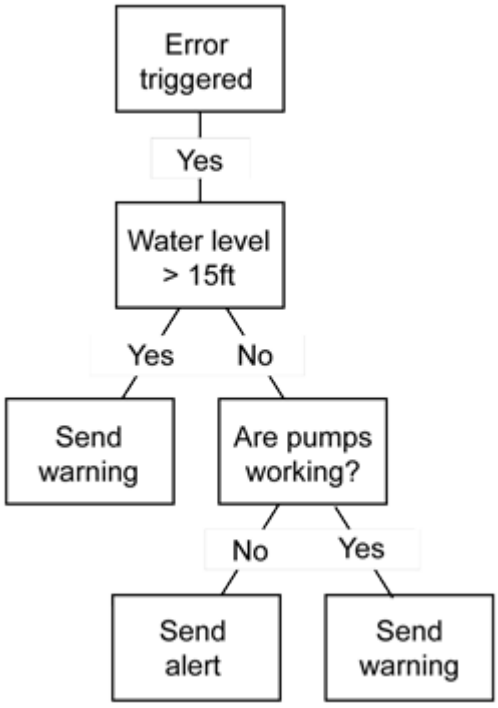
authorization, for example, you might combine decision tree analysis of individual past transactions with classification and historical pattern matches to identify whether a transaction is fraudulent. Making a match between the purchase of flights to the US and transactions in the US, it is likely that the transaction is valid.

Sequential patterns

Often used over longer-term data, sequential patterns are a useful method for identifying trends, or regular occurrences of similar events. For example, with customer data you can identify that customers buy a particular collection of products together at different times of the year. In a shopping basket application, you can use this information to automatically suggest that certain items be added to a basket based on their frequency and past purchasing history.

Decision trees

Related to most of the other techniques (primarily classification and prediction), the decision tree can be used either as a part of the selection criteria, or to support the use and selection of specific data within the overall structure. Within the decision tree, you start with a simple question that has two (or sometimes more) answers. Each answer leads to a further question to help classify or identify the data so that it can be categorized and classify an incoming error condition.



Decision trees are often used with classification systems to attribute type information, and with predictive systems, where different predictions might be based on past historical experience that helps drive the structure of the decision tree and the output.

III. RESULTS AND DISCUSSION

Three key features of *k*-means which make it efficient are :

- Euclideandistance is used as a matric and variance is used as a measure of cluster scatter.
- The number of clusters *k* is an input parameter: an inappropriate choice of *k* may yield poor results. That is why, when performing k-means, it is important to run diagnostic checks for determining the number of clusters in data set.
- Convergence to a local minimum may produce counterintuitive results.

A key limitation of *k*-means is its cluster model. The concept is based on spherical clusters that are separable in a way so that the mean value converges towards the cluster center. The clusters are expected to be of similar size, so that the assignment to the nearest cluster center is the correct assignment. When for example applying *k*-means with a value of onto the well-known Irisflowerdataset, the result often fails to separate the three Iris species contained in the data set. With, the two visible clusters (one containing two species) will be discovered, whereas with one of the two clusters will be split into two even parts. In fact, is more appropriate for this data set, despite the data set containing 3 *classes*. As with any other clustering algorithm, the *k*-means result relies on the data set to satisfy the assumptions made by the clustering algorithms. It works well on some data sets, while failing on others.

Decision treelearning uses a decisiontree (as a predictivemodel) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modeling approaches used in statistics, datamining and machinelearning. Tree models where the target variable can take a discrete set of values are called classificationtrees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically realnumbers) are called regressiontrees.In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decisionmaking. In datamining, a decision tree describes data (but the resulting classification tree can be an input for decisionmaking). This page deals with decision trees in datamining.

Predictive analytics encompasses a variety of statistical techniques from predictive modeling, machine learning, and

 ISSN: 2454-4248
397 – 400

data mining that analyze current and historical facts to make predictions about future or otherwise unknown events.In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decisionmaking for candidate transactions.The defining functional effect of these technical approaches is that predictive analytics provides a predictive score (probability) for each individual (customer, employee, healthcare patient, product SKU, vehicle, component, machine, or other organizational unit) in order to determine, inform, or influence organizational processes that pertain across large numbers of individuals, such as in marketing, credit risk assessment, fraud detection, manufacturing, healthcare, and government operations including law enforcement

IV. CONCLUSION

Data mining is more than running some complex queries on the data you stored in your database. You must work with your data, reformat it, or restructure it, regardless of whether you are using SQL, document-based databases such as Hadoop, or simple flat files. Identifying the format of the information that you need is based upon the technique and the analysis that you want to do. After you have the information in the format you need, you can apply the different techniques (individually or together) regardless of the required underlying data structure or data set.

REFERENCES

[1] Xue Li, , Vasu D. Chakravarthy, , Bin Wang, and Zhiqiang Wu, “Spreading Code Design of Adaptive Non-Contiguous SOFDM for Dynamic Spectrum Access” in IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, VOL. 5, NO. 1, FEBRUARY 2011

[2] J. D. Poston and W. D. Horne, “Discontiguous OFDM considerations for dynamic spectrum access in idel TV channels,” in Proc. IEEE DySPAN, 2005.

[3] R. Rajbanshi, Q. Chen, A.Wyglinski, G. Minden, and J. Evans, “Quantitative comparison of agile modulation technique for cognitive radio tranceivers,” in Proc. IEEE CCNC, Jan. 2007, pp. 1144–1148.

[4] V. Chakravarthy, X. Li, Z. Wu, M. Temple, and F. Garber, “Novel overlay/underlay cognitive radio waveforms using SD-SMSE framework to enhance spectrum efficiency—Part I,” IEEE Trans. Commun., vol. 57, no. 12, pp. 3794–3804, Dec. 2009.

[5] V. Chakravarthy, Z. Wu, A. Shaw, M. Temple, R. Kannan, and F. Garber, “A general overlay/underlay analytic expression for cognitive radio waveforms,” in Proc. Int. Waveform Diversity Design Conf., 2007.

[6] V. Chakravarthy, Z. Wu, M. Temple, F. Garber, and X. Li, “Cognitive radio centric overlay-underlay waveform,” in Proc. 3rd IEEE Symp. New Frontiers Dynamic Spectrum Access Netw., 2008, pp. 1–10.

[7] X. Li, R. Zhou, V. Chakravarthy, and Z. Wu, “Intercarrier interference immune single carrier OFDM via magnitude shift keying modulation,” in Proc. IEEE Global Telecomm. Conf. GLOBECOM , Dec. 2009, pp. 1–6.

[8] Parsaee, G.; Yarali, A., "OFDMA for the 4th generation cellular networks" in Proc. IEEE Electrical and Computer Engineering, Vol.4, pp. 2325 - 2330, May 2004.

[9] 3GPP R1-050971,"R1-050971 Single Carrier Uplink Options for EUTRA: IFDMA/DFT-SOFDM Discussion and Initial Performance Results ",http://www.3GPP.org,Aug 2005

[10] IEEE P802.16e/D12,'Draft IEEE Standard for Local and metropolitan area networks-- Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems', October 2005

[11] 3GPP RP-040461, Study Item: Evolved UTRA and UTRAN, December 200

[12] R. Mirghani, and M. Ghavami, “Comparison between Wavelet-based and Fourier-based Multicarrier UWB Systems”, IET Communications, Vol. 2, Issue 2, pp. 353-358, 2008.

[13] R. Dilmirghani, M. Ghavami, “Wavelet Vs Fourier Based UWB Systems”, 18th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, pp.1-5, Sep. 2007.

[14] M. Weeks, Digital Signal Processing Using Matlab and Wavelets, Infinity Science Press LLC, 2007.

[15] S. R. Baig, F. U. Rehman, and M. J. Mughal, “Performance Comparison of DFT, Discrete Wavelet Packet and Wavelet Transforms in an OFDM Transceiver for Multipath Fading Channel,”,9th IEEE International Multitopic Conference, pp. 1-6, Dec. 2005.

[16] N. Ahmed, Joint Detection Strategies for Orthogonal Frequency Division Multiplexing, Dissertation for Master of Science, Rice University, Houston, Texas. pp. 1-51, Apr. 2000.