Data Mining Techniques for Text Mining Using Clustering and Classifiers

D.Nalini Kumari [1], G.M.Padmaja [2],
Assistant Professor [1],[2],
Department of CSE
Mail Id : nalinimtech10@gmail.com, Mail id : padmaja.gmp@gmail.com,

Abstract—

In the past two decades, databases and data mining methods have developed into a powerful tool for dealing with massive amounts of data. Association rule mining, clustering, classification, and other methods are all examples of data mining processes. Clustering is a well-known and often used data mining method. Clustering is an effective method used in machine learning to group items that have common characteristics. Its efficiency and ability to unearth previously undiscovered information in massive, complicated datasets have earned it widespread renown. The quality of a cluster, however, is subjective and relies on the domain in which it is used. The primary objective is to improve upon traditional clustering algorithms like k-means and EM (Expectation Maximization) in order to provide more accurate groups. The healthcare industry is a knowledge-rich environment, therefore it makes sense that data mining expertise should also increase to meet rising demand in this area. Tacit healthcare knowledge is an essential part of the healthcare delivery system, but it has been underutilized because of its intangible character. There are a lot of algorithms that attempt to solve these issues, but they often fail to do so precisely. Patient-reported outcomes (PROs) have been more common in clinical research due to their significance in assessing medicines and shaping treatment strategies. Health care data is growing at an unprecedented pace, and much of it is being gathered, loaded, stored, and accessible over the internet.

I. INTRODUCTION

Humans are unable to manage the vast amounts of data generated in the medical industry. data, more and more focus has shifted to the processing of such data. In order to meet all of these requirements, it is crucial to develop better clustering methods. Clustering is an unsupervised classification method that may be used to any dataset to create distinct groups of items. Class objects should have common characteristics. Only the characteristics of the objects in the data set are visible at first glance; no information about the classes or the number of classes is provided. Our primary scientific contribution is a novel method of using distance measures with the K-means Algorithm.

1.1 Mining for Data

Data mining, knowledge extraction, information discovery, data archeology, and data pattern processing are just a few of the many titles that have been used to describe the process of identifying meaningful patterns in large datasets. Data mining is a prominent word in the fields of statistics and database management. The first KDD workshop was held in 1989 (Piatetsky-Shapiro 1991), and since then, the phrase "knowledge discovery in databases" (KDD) has become commonplace in the domains of artificial intelligence and machine learning. (Usama Fayyad) By definition, data mining is a cutting-edge technique used to sift through large amounts of data in search of the most relevant information. There is a pressing need for computational technologies that can meet the problems presented by the plethora of fresh data sets being gathered across many disciplines.

Pattern recognition and data mining have a long and storied history. Bayes' theorem (1700s) and regression analysis (1800s) have been utilized before [M. S. V. K. 2006]. Using the ever-increasing processing power of computers, we create a vital resource for manipulating data in the area of information technology. For instance, it has features like the capacity to deal with ever-increasing dataset sizes and complexity. The capacity to gather, store, and manipulate data has improved thanks to advancements in computer science, and there is a pressing need to improve automated data processing. Wikipedia lists the development of the neural network, cluster analysis, the genetic algorithm (1950s), decision trees (1960s), and support vector machines (1990s) as examples of such seminal achievements.

Information extraction from ever-increasing data quantities is driving the development of the data-mining sector. It looks for hidden insights in the data that standard analysis tools can't unearth. Data mining is, as we indicated before, an essential aspect of KDD; in our perspective, KDD describes the larger process of extracting actionable insights from data, while data mining is a subset of that process. Figure 1.1 illustrates the KDD function: transforming raw data into useful knowledge.

From initial data collection to the final data mining findings, quite a few transformational processes are involved in this process. [M. S. V. K. 2006]

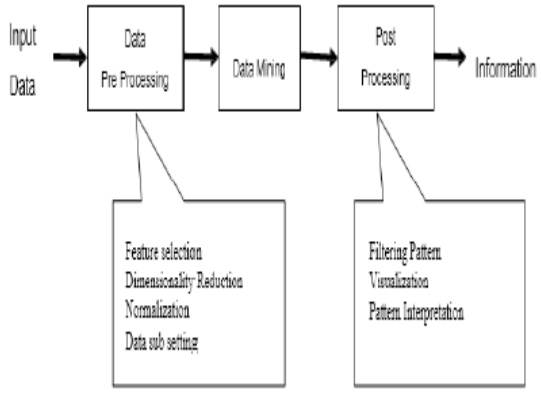

Figure 1.1: A schematic of the KDD procedure

Input Data may be located in a centralized Data Repository or across numerous locations, and its storage format can range from a simple text file or spreadsheet to a more complex relational database. The Pre-processing Phase has been completed to prepare the raw data for further analysis. This phase entails combining information from several sources, eliminating irrelevant details, and keeping track of what's important for the data mining process. Due to the wide variety of data, this process takes the longest and requires the greatest effort.

II. RESEARCH METHODOLOGY

Although data mining is still in its infancy as a field of endeavor, numerous organizations across many sectors, including but not limited to the financial sector, the healthcare industry, the manufacturing sector, the transportation sector, and the aerospace sector, are already making use of data mining tools and techniques to profit from accumulated data. Using pattern recognition technology, statistical, and mathematical methodologies, data mining aids in the analysis of relationships, trends, patterns, exceptions, and unexpected data that could otherwise go missed. Better business decisions may be made with the help of data mining, which involves uncovering patterns and correlations in large amounts of data. It may aid in creating effective marketing strategies and predicting client loyalty.

Some concrete applications of data mining are: [D.Alexander,"Datamining]

Customers that purchase similar items from your organization have certain traits, which may be used for market segmentation.

Predicting which consumers may defect to a rival business is called "customer churn."

• Fraud detection identifies potentially fraudulent transactions.

For the direct marketing question:

The best response rate may be achieved by include prospects on a mailing list.

• Interactive advertising - guess what each website visitor wants to see before they arrive.

To learn what people often buy together, like beer and diapers, a "market basket analysis" might be performed.

Using a trend analysis, describe how your usual consumers this month vary from those of previous.

Projects in Data Mining 2.1

There are two main types of data mining projects:

Exercises in foresight: The purpose of this exercise is to determine the likely worth of a single quality given the values of many others. Explanatory or independent variables are those utilized to make a prediction, whereas the value being predicted is called the target or dependent variable.

Job of description:

The goal of this exercise is to infer latent connections between variables. Independent values are required for this data mining activity, which sometimes necessitates additional processing steps before conclusions can be drawn.

Figure 2.1 depicts four of the most popular types of data mining activities; the next section elaborates on cluster analysis.

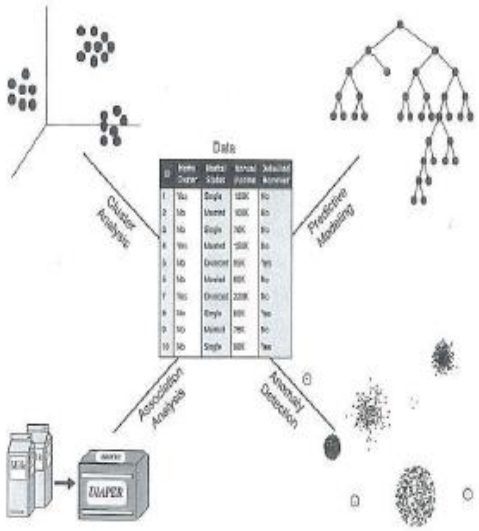

Figure 2.1: Four fundamental data mining operations

The six essential data mining tasks are as follows:

Classification is used for discrete target variables, whereas Regression is utilized for continuous ones in predictive modeling tasks. In both cases, you need to create a model that can predict values with as little error as possible. Emails may be categorized as authentic or spam, and web users can be predicted as to whether or not they will make a purchase at an online bookstore. The price, however, is a continues-value characteristic, making price forecasting a Regression problem. Dependency modeling, also known as Association Rule Learning, is a technique for describing linked aspects in data and discovering hidden associations. Association analysis may reveal, for instance, pairs of websites that are often visited together.
In the field of Anomaly Detection (also known as Outlier, Change, and Deviation Detection), students learn to spot unusual data that either causes mistakes or might be of interest and warrants further inquiry. Clustering is another category; it's the process of finding groupings and structures inside data that are "similar" or "dissimilar" in some way, without resorting to previously-known structures within the data (this job will be explored in more depth in the subsequent section of the report).Finally, the Summarization class aims to provide reports and visual representations of data in a condensed format. (Usama Fayyad)

III. CONCLUSION

This research looks at the previous 12 years (2004-2015) of design research and gives a bibliometric, network-theoretic, and text-based examination of the field. Our work is the first comprehensive investigation into employing text mining methods to monitor developments in the field of design research. It also demonstrates that the techniques established are generalizable and may be used to organize the data and information from a wide range of scientific disciplines. In addition, the text mining methods utilized here might provide scholars with a holistic view of the expertise in a particular topic that is otherwise buried in a mountain of scholarly literature. The clustering technique offers a more comprehensive perspective of a field's underlying architectural framework. In addition, social network analysis delves deeper into these core themes to provide academics with a more complete picture of a field's progress. For future research, we plan to implement the author-topic model, a probabilistic model to link authors to observed words in the scientific literature of the research fields (a case study of design research), which will serve as a general framework for author-topic relationship exploration, discovery, and query-answering.

REFERENCES

1. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. J. Mach. Learn. Res. 2003, 3, 993–1022.

2. Hartigan, J.A.; Wong, M.A. A k-means clustering algorithm. Appl. Stat. 1979, 28, 100–108. 11. Heinrich, G. Parameter Estimation for Text Analysis. Available online:

3. Singh, V.K.; Uddin, A.; Pinto, D. Computer science research: The top 100 institutions in india and in the world. Scientometrics 2015, 104, 529–553. [CrossRef]

4. Sitarz, R. Identification of Research Trends in the Field of Separation Processes. Application of Epidemiological Model, Citation Analysis, Text Mining, and Technical Analysis of the Financial Markets. Available online: http://www.doria.fi/bitstream/handle/10024/93330/isbn9789522654847.pdf?sequence=2 (accessed on 14 April 2017).

5. Saffer, D. Designing for Interaction: Creating Innovative Applications and Devices; New

Riders: Berkley, CA, USA, 2010. 15. Garrett, J.J. Elements of User Experience: User-Centered Design for the Web and Beyond; New Riders: Berkley, CA, USA, 2010.

6. Ferguson, S.A.; Allread, W.G.; Le, P.; Rose, J.; Marras, W.S. Shoulder muscle fatigue during repetitive tasks as measured by electromyography and near-infrared spectroscopy. Hum. Factors J. Hum. Factors Ergon. Soc. 2013, 55, 1077–1087. [CrossRef] [PubMed]

7. Jeon, M.; Walker, B.N.; Gable, T.M. The effects of social interactions with in-vehicle agents on a driver's anger level, driving performance, situation awareness, and perceived workload. Appl. Ergon. 2015, 50, 185–199. [CrossRef] [PubMed]