Business Intelligence Using Data Mining Techniques on Very Large Datasets

Arti J. Ugale[1], P. S. Mohod[2]

[1]Department of Computer Science and Engineering, RTMNU

[2]Professor, Department of Computer Science and Engineering, RTMNU

Abstract: ***Business Intelligence (BI) is a concept of applying a set of technologies to convert data into meaningful information. BI methods include information retrieval, data mining, statistical analysis as well as data visualization. Large amounts of data originating in different formats and from different sources can be consolidated and converted to key business knowledge. Data mining is used to search for patterns and correlations within a database of information. Business intelligence (BI) focuses on detail integration and organization. DM aids BI's objectives.DM and BI work together to process data and analyze it in a way that eases the workload for the users and aids with the understanding of the materials/findings. This is accomplished through recognizing relationships in the data and identifying opportunities and risks of the company. It also allows users to manipulate the data to fulfil their specific user-oriented objectives. Data mining is the process of searching through data using various algorithms to discover patterns and correlations within a database of information. Business intelligence, on the other hand, focuses more on data integration and organization. It will combine data analyse to help managers make operational, tactical, or strategic business decisions. Data mining can be used to aid the objectives of a business intelligence system.Classification and patterns extraction from customer data is very important for business support and decision making. Timely identification of newly emerging trends is very important in business process. Large companies are having huge volume of data but starving for knowledge. To overcome the organization current issue, the new breed of technique is required that has intelligence and capability to solve the knowledge scarcity and the technique is called Data mining. The objectives of this paper are to identify the high-profit, high-value and low-risk customers by one of the data mining technique – customer clustering. The paper explores the concepts of BI, its components, emergence of BI, benefits of BI, factors influencing BI, technology requirements, designing and implementing business intelligence, and various BI techniques.***

Keywords: Data Mining, Business Intelligence, Distributed algorithm, Clustering, Content Based Indexing.

1. Introduction

Using technology to gain an edge in business is not a new idea. Whenever there is something new, entrepreneurs will be quick to try to find an application for it in the business world to make money. Data mining (DM) and business intelligence (BI) are among the information technology applications that have business value. This paper will first outline what data mining and business intelligence are, then move on to practical usages in various business contexts. It will then proceed to a section dealing with how C-Suite executives, like the CFO, CIO, etc., will handle the choice of whether to implement a system and how to go about doing it. Suggestions as to which industries are best suited for this technology are also given. Finally, there is a section on how DM and BI will affect the accounting profession. Data mining is the process of searching through data using various algorithms to discover patterns and correlations within a database of information. Business intelligence, on the other hand, focuses more on data integration and organization. It will combine data analyse to help managers make operational, tactical, or strategic business decisions. Data mining can be used to aid the objectives of business intelligence system. Business Intelligence could be a idea of applying a group of technologies to convert information into meaning data. Bismuth ways embody data retrieval, data processing, applied math analysis yet as information visual image. Giant amounts of knowledge| of information originating completely different in several numerous formats and from different sources may be consolidated and regenerate to key business knowledge. Presents a general read on however information square measure remodelled to business intelligence. The method involves each business consultants and technical consultants. It converts an outsized scale of information to meaning outcomes therefore on offer decision-making support to finish users. Business intelligence (BI) has two basic different meanings related to the use of the term intelligence. The primary, less frequently, is the human intelligence capacity applied in business affairs/activities. Intelligence of Business is a new field of the investigation of the application of human cognitive faculties and artificial intelligence technologies to the management and decision support in different business problems. The second relates to the intelligence as information valued for its currency and relevance. It is expert information, knowledge and technologies efficient in the management of organizational and individual business. Therefore, in this sense, business intelligence is a broad category of applications and technologies for gathering, providing access to, and analyzing data for the purpose of helping enterprise users make better business decisions. The term implies having a comprehensive knowledge of all of the factors that affect the business. It is imperative that firms have an in depth knowledge about factors such as the customers, competitors, business partners, economic environment, and internal operations to make effective and good quality business decisions. Business intelligence enables firms to make these kinds of decisions. The paper explores the concepts of BI, its components, emergence of BI, benefits of BI, factors influencing BI, technology requirements, designing and implementing business intelligence, cultural imperatives, and various BI techniques. The paper would be useful for budding researchers in the field of BI to understand the basic concepts.

2. A Brief Literature Survey

In paper [1], In this paper, we presented a two-stage method for creating accurate classifiers for DNA sequences with interesting and comprehensible features.Data mining is the extraction of hidden predictive information from large databases and it is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses.In paper [2], different feature selection techniques for classification: Big Data concerns large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data is now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. In paper [3], this paper presents an application of business intelligence (BI) for electricity management systems in the context of the Smart Grid domain. This distribution is the basis for performing on an upper level all the business processes for managing the energy demand and other customer services in a Smart Grid. We study here the problem of secure mining of association rules in horizontally partitioned databases. In that setting, there are several sites (or players) that hold homogeneous databases,i.e., databases that share the same schema but hold information on different entities. In paper [4] this paper proposed a data mining methodology called business intelligence driven data mining. It combines the knowledge driven data mining and method driven data minig and fills the gap between business intelligence knowledge and existence various data mining methods in e-Business. It setup a four layer frame layer. outlier detection is the data mining task whose goal is to isolate the observations which are considerably dissimilar from the remaining data. This task has practical applications in several domains such as fraud detection, intrusion detection, data cleaning, medical diagnosis, and many others.In paper [5] this paper presents an application of business intelligence for electricity management systems in the context of the Smart Grid domain. This distribution is the basis for performing on an upper level all the business processes for managing the energy demand and other customer services.In paper[6] proposed a data mining methodology called business intelligence driven data mining. It combines the knowledge driven data mining and method driven data minig and fills the gap between business intelligence knowledge and existence various data mining methods in e-Business.

3. Process of Execution

Two data-mining algorithms, Distributed Datawarehouse algorithm (DDM), Content Based Indexing (CBI),clustering and classification were used consists of a group of interconnected a set data. This methodology are concerned with efficient storage and retrieval of records. Knowledge Layer: This layer is on top interfacing with end-users. It is referered as a knowledge or business decision-making support. This layer harvests business intelligence from patterns that are derived from business data. Knowledge of customers shopping behaviour. Method Layer: This layer contains data minig algorithms, which is use to transfer data into some meaningful expression. Data Layer: This layer is responsible for providing data source of knowledge discovery. The data source has been preprocessed, which means transforming the raw data to cleaned data.

- To retrieve and analyze the data
- To extract, transform and load data
- To manage data dictionary

4. Methodology

A. Data Classification

The data used for this work was collected from different showrooms. Classification may refer to categorization, the process in which ideas and objects are recognized, differentiated, and understood. In many classification applications, Support Vector Machines (SVMs) have proven to be highly performing and easy to handle classifiers with very good generalization abilities. A large classification problem can be split into mainly easy and only a few hard subproblems. On standard benchmark datasets, this approach achieved great speedups while suffering only slightly in terms of classification accuracy and generalization ability.

B. Clustering

Cluster analysis is an exploratory data analysis tool for solving classification problems. Its object is to sort cases (people, things, events, etc) into groups, or clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters. Each cluster thus describes, in terms of the data collected, the class to which its members belong; and this description may be abstracted through use from the particular to the general class or type. Clustering is the task of grouping a set of objects in such a way that objects in the same group called a **cluster** are more similar to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and informatics. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. Customer clustering is the most important data mining methodologies used in marketing and customer relationship management .

C. Distributed Data warehouse Algorithm

Identifying how the data is distributed is the first step in developing a distributed data mining algorithm.DDM is develops for relational data model. The DDM algorithm is designed upon the potential parallelism they can apply on the given data. In the field of data management, data classification as a part of Information Lifecycle

Management process can be defined as a tool for categorization of data to enable/help organization to effectively answer following questions:

- What data types are available?
- Where are certain data located?
- What access levels are implemented?

It is worth to observe that several mining algorithms deal with distributed data set by computing local models which are aggregated in a general model as a final step in the supervisor node. This algorithm is different, since it computes the true global model through iterations where only selected global data and all the local data are involved. Data warehouses rule square measure designed to assist you analyze information. for instance, to find out a lot of regarding your company's sales information, you'll build a warehouse that concentrates on sales.

D. Content Based Indexing Algorithm

In this when an algorithm is develop on dataset for index formation and more accuracy. By content based indexing algorithm exact data find out to process the dataset by indexing the clusters. These methodologies are concerned with efficient storage and retrieval of records. The current technology of text-based indexing and retrieval implemented for relational databases does not provide practical solutions for this problem of managing huge multimedia repositories. Most of the commercially available multimedia indexing and search systems index the media based on keyword annotations and use standard text based indexing and retrieval mechanisms to store and retrieve multimedia data. There are often many limitations with this method of keywords based indexing and retrieval especially in the context of multimedia databases. First, it is often difficult to describe with human languages the content of a multimedia object, for example an image having complicated texture patterns. Second, manual annotation of text phrases for a large database is prohibitively laborious in terms of time and effort. Third, since users may have different interests in the same multimedia object, it is difficult to describe it with a complete set of key words. Finally, even if all relevant object characteristics are annotated, difficulty may still arise due to the use of different indexing languages or vocabularies by different users. In content-based retrieval, manual annotation of visual media is avoided and indexing and retrieval is instead performed on the basis of media content itself. There have been extensive studies on the design of automatic *content-based indexing and retrieval* (CBIR) systems.

E: Project Flow Diagram

The overall project is done with the help of Algorithm in the stages of following data flow diagram:

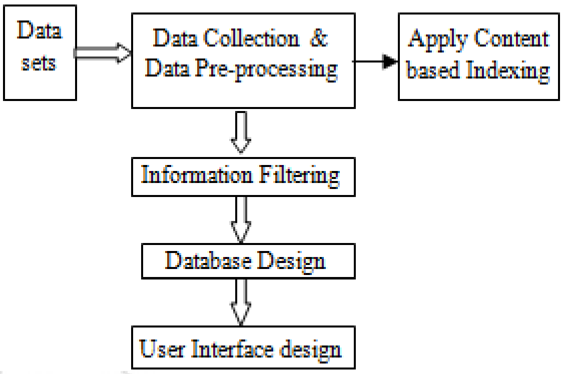


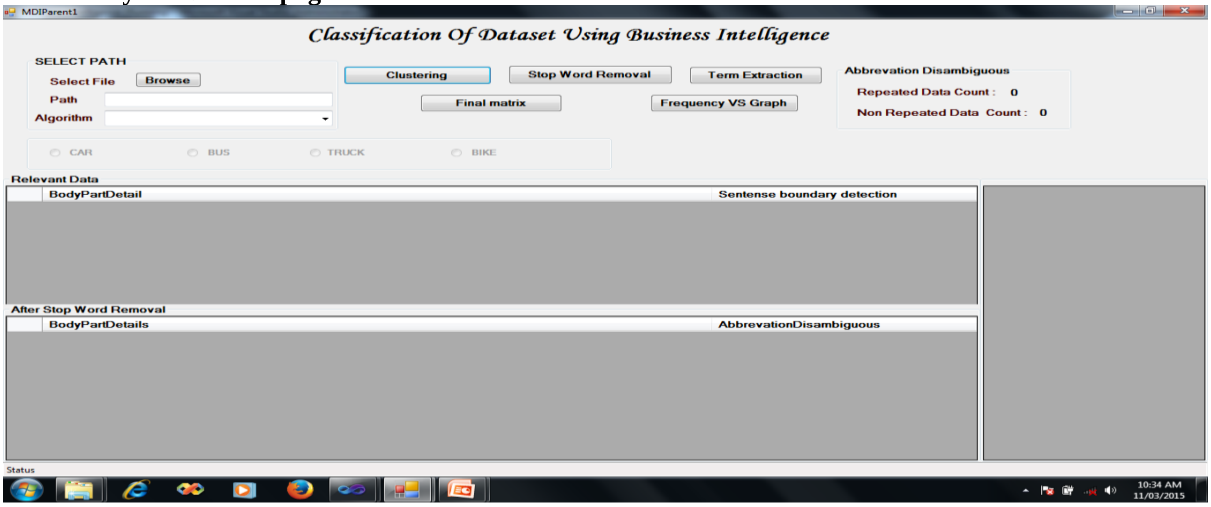
Figure: Work Flow

In this approach project is completed in four stages as shown in above figure i.e data collection and data pre processing, data filtering, database design and finally user interface design. The primary activities include gathering, preparing and analyzing data. The data itself must be of high quality. The various sources of data is collected, transformed, cleansed, loaded and stored in a warehouse. The relevant data is for a specific business area that is extracted from the data warehouse. Analytical Processing provides multidimensional, summarized views of business data and is used for reporting, analysis, modeling and planning for optimizing the business.

E. Result

On available input data sets we apply Content Based Indexing Algorithm a robust classification of datasets and indexes it with proper way by using the current state regarding how data is collected, analyzed, and disseminated including what infrastructure, tools and support exist.

E.1 GUI for the system: Home page

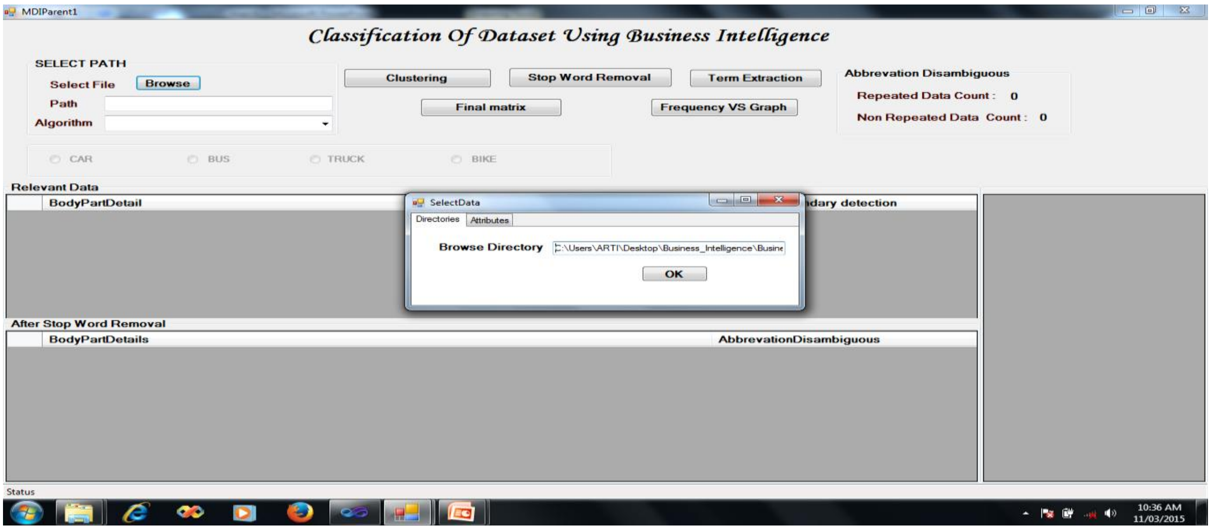


Classification Of Dataset Using Business Intelligence

Repeated Data Count : 0

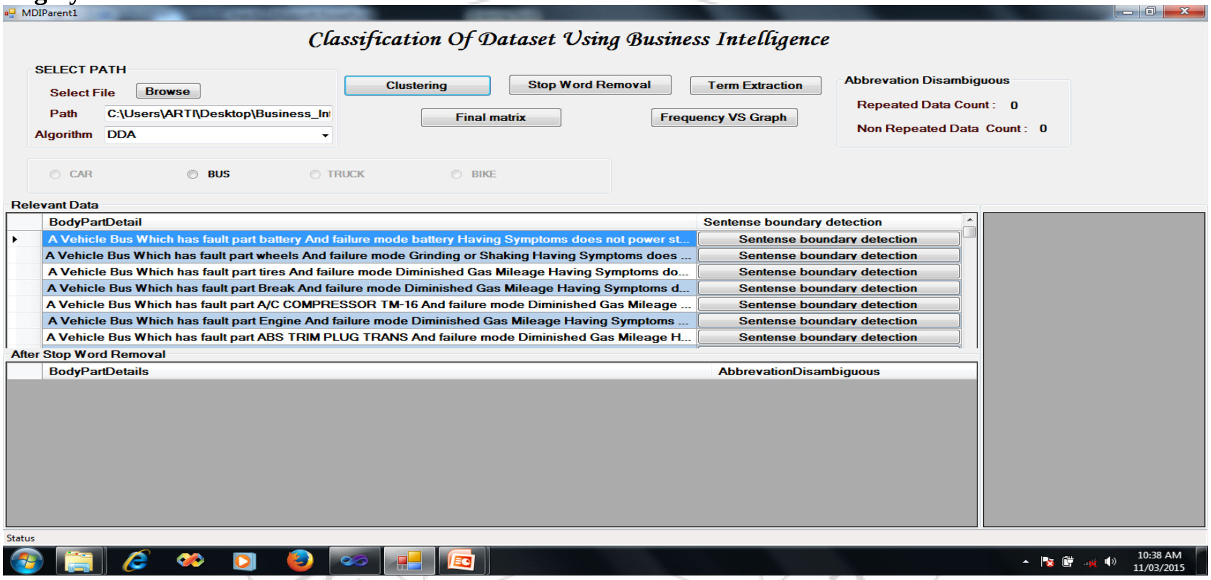
Non Repeated Data Count : 0

Data Selection:



Browse Directory: C:\Users\ARTI\Desktop\Business_Intelligence\Busine

Clustering by DDA:



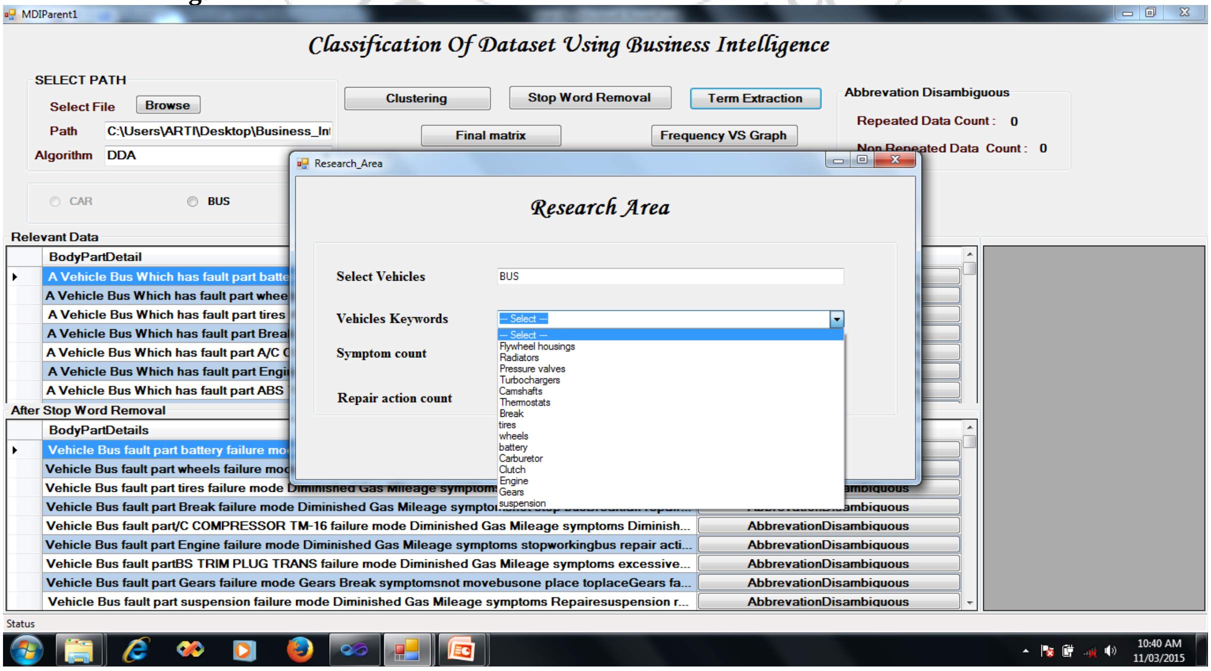
Path: C:\Users\ARTI\Desktop\Business_Int

Algorithm: DDA

○ CAR ● BUS ○ TRUCK ○ BIKE

BodyPartDetail	Sentense boundary detection
A Vehicle Bus Which has fault part battery And failure mode battery Having Symptoms does not power st...	Sentense boundary detection
A Vehicle Bus Which has fault part wheels And failure mode Grinding or Shaking Having Symptoms does ...	Sentense boundary detection
A Vehicle Bus Which has fault part tires And failure mode Diminished Gas Mileage Having Symptoms do...	Sentense boundary detection
A Vehicle Bus Which has fault part Break And failure mode Diminished Gas Mileage Having Symptoms d...	Sentense boundary detection
A Vehicle Bus Which has fault part A/C COMPRESSOR TM-16 And failure mode Diminished Gas Mileage ...	Sentense boundary detection
A Vehicle Bus Which has fault part Engine And failure mode Diminished Gas Mileage Having Symptoms ...	Sentense boundary detection
A Vehicle Bus Which has fault part ABS TRIM PLUG TRANS And failure mode Diminished Gas Mileage H...	Sentense boundary detection

Content Based Indexing:



Path: C:\Users\ARTI\Desktop\Business_Int

Algorithm: DDA

Select Vehicles: BUS

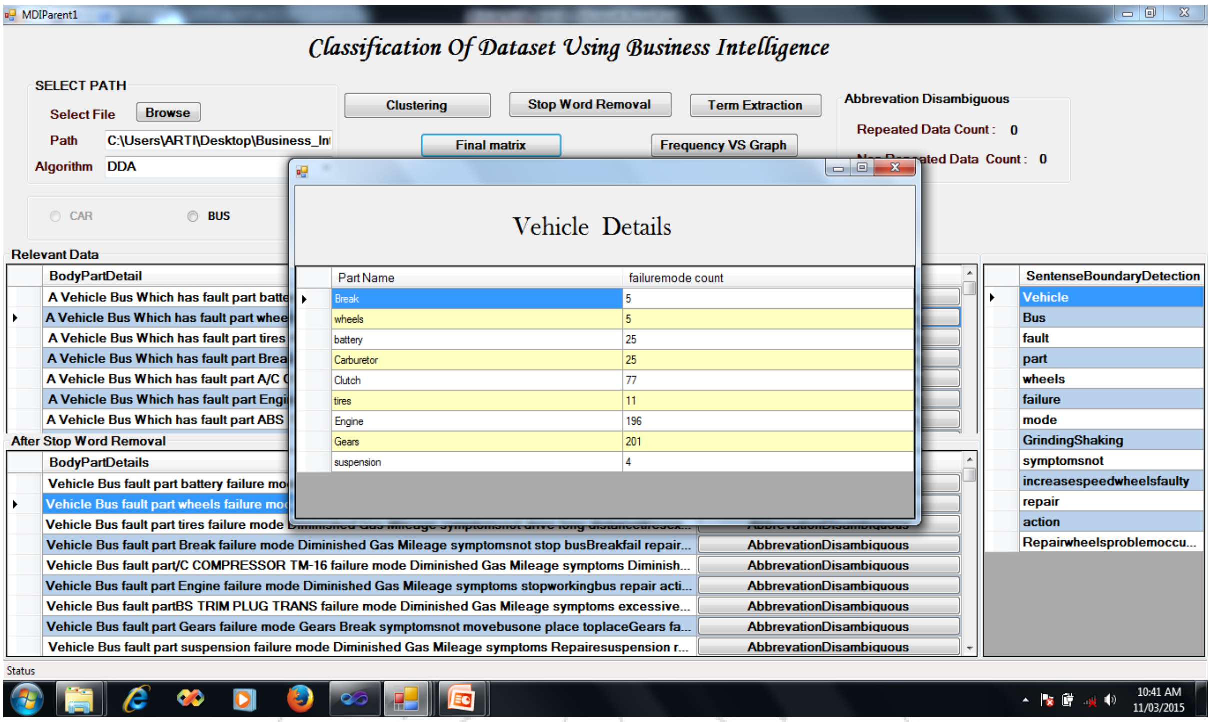
Vehicles Keywords: -- Select --

-- Select --
Flywheel housings
Radiators
Pressure valves
Turbochargers
Camshafts
Thermostats
Break
tires
wheels
battery
Carburetor
Clutch
Engine
Gears
suspension

BodyPartDetail
A Vehicle Bus Which has fault part batte
A Vehicle Bus Which has fault part whee
A Vehicle Bus Which has fault part tires
A Vehicle Bus Which has fault part Brea
A Vehicle Bus Which has fault part A/C C
A Vehicle Bus Which has fault part Engi
A Vehicle Bus Which has fault part ABS

BodyPartDetails	
Vehicle Bus fault part battery failure mo	
Vehicle Bus fault part wheels failure mod	
Vehicle Bus fault part tires failure mode Diminished Gas Mileage symptoms	...ambiguous
Vehicle Bus fault part Break failure mode Diminished Gas Mileage sympto...	...ambiguous
Vehicle Bus fault part/C COMPRESSOR TM-16 failure mode Diminished Gas Mileage symptoms Diminish...	AbbrevationDisambiguous
Vehicle Bus fault part Engine failure mode Diminished Gas Mileage symptoms stopworkingbus repair acti...	AbbrevationDisambiguous
Vehicle Bus fault partBS TRIM PLUG TRANS failure mode Diminished Gas Mileage symptoms excessive...	AbbrevationDisambiguous
Vehicle Bus fault part Gears failure mode Gears Break symptomsnot movebusone place toplaceGears fa...	AbbrevationDisambiguous
Vehicle Bus fault part suspension failure mode Diminished Gas Mileage symptoms Repairesuspension r...	AbbrevationDisambiguous

Final Matrix:



Classification Of Dataset Using Business Intelligence

SELECT PATH
Select File
Path C:\Users\ARTI\Desktop\Business_Int
Algorithm DDA
○ CAR ○ BUS

Abbrevation Disambiguous
Repeated Data Count : 0
Non Repeated Data Count : 0

Vehicle Details

Part Name	failuremode count
Break	5
wheels	5
battery	25
Carburetor	25
Clutch	77
tires	11
Engine	196
Gears	201
suspension	4

Relevant Data

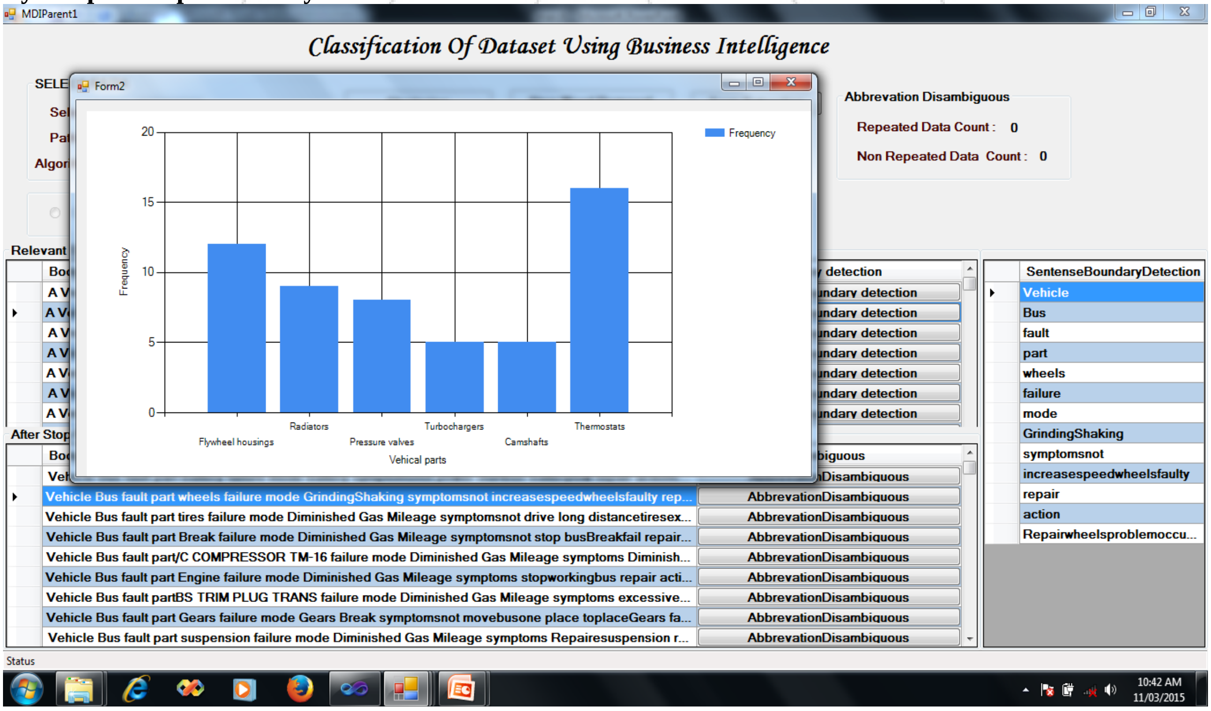
BodyPartDetail
A Vehicle Bus Which has fault part batte
A Vehicle Bus Which has fault part whee
A Vehicle Bus Which has fault part tires
A Vehicle Bus Which has fault part Brea
A Vehicle Bus Which has fault part A/C C
A Vehicle Bus Which has fault part Engi
A Vehicle Bus Which has fault part ABS

After Stop Word Removal

BodyPartDetails	
Vehicle Bus fault part battery failure mo	
Vehicle Bus fault part wheels failure mo	
Vehicle Bus fault part tires failure mode	
Vehicle Bus fault part Break failure mode Diminished Gas Mileage symptomsnot stop busBreakfail repair...	AbbrevationDisambiquous
Vehicle Bus fault part/C COMPRESSOR TM-16 failure mode Diminished Gas Mileage symptoms Diminish...	AbbrevationDisambiquous
Vehicle Bus fault part Engine failure mode Diminished Gas Mileage symptoms stopworkingbus repair acti...	AbbrevationDisambiquous
Vehicle Bus fault partBS TRIM PLUG TRANS failure mode Diminished Gas Mileage symptoms excessive...	AbbrevationDisambiquous
Vehicle Bus fault part Gears failure mode Gears Break symptomsnot movebusone place toplaceGears fa...	AbbrevationDisambiquous
Vehicle Bus fault part suspension failure mode Diminished Gas Mileage symptoms Repairesuspension r...	AbbrevationDisambiquous

SentenseBoundaryDetection
Vehicle
Bus
fault
part
wheels
failure
mode
GrindingShaking
symptomsnot
increasespeedwheelsfaulty
repair
action
Repairwheelsproblemoccu...

Freuency Graph: Output of the system



Classification Of Dataset Using Business Intelligence

Abbrevation Disambiguous
Repeated Data Count : 0
Non Repeated Data Count : 0

BodyPartDetails	
Vehicle Bus fault part wheels failure mode GrindingShaking symptomsnot increasespeedwheelsfaulty rep...	AbbrevationDisambiquous
Vehicle Bus fault part tires failure mode Diminished Gas Mileage symptomsnot drive long distancetiresex...	AbbrevationDisambiquous
Vehicle Bus fault part Break failure mode Diminished Gas Mileage symptomsnot stop busBreakfail repair...	AbbrevationDisambiquous
Vehicle Bus fault part/C COMPRESSOR TM-16 failure mode Diminished Gas Mileage symptoms Diminish...	AbbrevationDisambiquous
Vehicle Bus fault part Engine failure mode Diminished Gas Mileage symptoms stopworkingbus repair acti...	AbbrevationDisambiquous
Vehicle Bus fault partBS TRIM PLUG TRANS failure mode Diminished Gas Mileage symptoms excessive...	AbbrevationDisambiquous
Vehicle Bus fault part Gears failure mode Gears Break symptomsnot movebusone place toplaceGears fa...	AbbrevationDisambiquous
Vehicle Bus fault part suspension failure mode Diminished Gas Mileage symptoms Repairesuspension r...	AbbrevationDisambiquous

SentenseBoundaryDetection
Vehicle
Bus
fault
part
wheels
failure
mode
GrindingShaking
symptomsnot
increasespeedwheelsfaulty
repair
action
Repairwheelsproblemoccu...

5. Conclusion

In this paper, Business intelligence is useful to obtain some guided data mining methods by identifying the related services. The core idea and information behind this architecture is to design a generic system that is flexible enough to suit to different requirement of the business. The business intelligence data of large enterprises can help to generate a very powerful knowledge base. We investigate the issue of classifying large dataset to mixed attributes data and present a constrained clustering algorithm as data redundancy is avoided,fast retrieval is possible faster decision-making and most accurate.

References

[1] Wendy Ashlock, *Student Member*, *IEEE*, and Suprakash Datta," Evolved Features for DNA Sequence Classification and Their Fitness Landscapes" IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 17, NO. 2, APRIL 2013

[2] S.C. Punitha, R. Jayasree, Dr. M. Punithavalli ," Partition Document Clustering using Ontology Approach" 2013 International Conference on Computer Communication and Informatics (*ICCCI* -2013), Jan. 04– 06, 2013, Coimbatore, INDIA

[3] Xindong Wu1,2, Xingquan Zhu3, Gong-Qing Wu2, Wei Ding4"Data mining with big data" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING.vol2.APRIL 2012.

[4] Angelina Espinoza, *Member*, *IEEE*, Yoseba Penya, *Senior Member*, *IEEE*, Juan Carlos Nieves,Mariano Ortega, *Member*, *IEEE*, Aitor Peña, and Daniel Rodríguez" Supporting Business Workflows in Smart Grids: An Intelligent Nodes-Based Approach"IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, 3 AUGUST 2013.

[5] Tamir tassa,"Secure Mining of Association Rules in Horizontally Distributed Databases",IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING

[6] Yang Hang, Simon Fong,"framework of business intelligence driven data mining for e-Business"2010fifth international join conference.

[7] FabrizioAngiulli, Senior Member, IEEE, Stefano Basta, Stefano Lodi, and Claudio Sartori, Distributed Strategies for Mining Outliers in Large Data Sets, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 7, JULY 2013.

[8] Mohammad Hassan Falakmasir, Shahrouz Moaven, Hassan Abolhassani, Jafar Habibi," Business Intelligence in E-Learning"