Dimension Debasing towards Minimal Search Space Utilization for Mining Patterns in Big Data

Dr. M. Naga Ratna
Dept. of Computer Science
JNTUH College of Engineering
Email: mratnajntu@jntuh.ac.in

Dara Karunya
Dept. of Computer Science
JNTUH College of Engineering
e-mail: karunya.dara222@gmail.com

Abstract **--**Data mining algorithms generally produce patterns which are interesting. Such patterns can be used by domain experts in order to produce business intelligence. However, most of the existing algoritms that can not properly work for uncertain data. Keeping uncertain data's characteristics in mind, it can be said that they do have more search space with existing algorithms. In this paper we proposed a method that can be used to reduce search space besides helping in producing patterns from uncertain data. The proposed method is based on MapReduce programming framework that works in distributed environment. The method essentially works on big data which is characterized by velocity, volume and variety. The proposed method also helps users to have constraints so as to produce high quality patterns. Such patterns can help in making well informed decisions. We built a prototype application that demonstrates the proof of concept. The empirical results are encouraging in mining uncertain big data in the presence of constraints.

Index Terms **–** *Big data, big data mining, uncertain data, frequent patterns, big data analytics*

*****

I. INTRODUCTION

Big data does mean huge amount of data that is measured generally in peta bytes [1]. In the industry, it is a hot topic or buzz word now as it can process huge amount of data that cannot be done in the traditional environments [2]. According to [3] Big Data mining attracted many researchers that make use of Hadoop as distributed programming framework. MapReduce is the new programming paradigm used for processing big data. This programming is done in distributed environment and it makes use of parallel processing. Over few years Hadoop became a reliable distributed programming framework. This framework is compatible with big data which is characterized by volume, velocity and variety. Hadoop is the framework that can handle such data which is static, dynamic, streaming and with various kinds. It also takes care of data correlation [4]. Big data processing has become an essential thing for enterprises as it can transform that data into economy of business. In other words big data mining can produce reliable business intelligence and its impact on the society is so high[5], [6] and[7]. Real value required by enterprises can be obtained from big data mining. To make the exercise fruitful it is important to make use of right kind of tools [8]. Three Vs are associated with big data that are Volume, Velocity and Variety[9]. Figure 1 conceptually shows this.

Fig. 1 – Big data is measured in Peta bytes or higher [1]

Big data can be transformed into big value so that organizations can benefit from that so as to make expert decisions [2]. The new programming paradigm used for big data processing is MapReduce which can process voluminous data in a comprehensive fashion [5]. With various applications that came into existence, Big Data mining became simple. The applications include GraphLab, Twitter Storm, Haloop, Twister, Spark and Apache Hama [10]. Solid State Drives (SSDs) and NAND flash memory are used for storing big data. The high speed processing of

 ISSN: 2321-8169
5090 - 5094

big data can increase ROI of companies [11]. Processing huge amount of data with latest cutting edge technologies can help companies to save money and time besides improving their financial status [12]. The big data mining and its results can be used various fields such as health care, agriculture, education and financial services to name few [13].

In this paper we focused on the minimum search space utilization for big data mining besides working on uncertain data. The remainder of the paper is structured as follows. Section 2 presents review of literature. Section 3 provides details of the proposed system. Section 4 presents experimental results while section 5 concludes the paper.

II. RELATED WORKS

Hadoop is a distributed file system that can be used to store and retrieve huge amount of data. The data is processed using a new programming paradigm known as MapReduce. Bu, Howe, and Ernst [14] studied and made a new variant of Hadoop. The intention is to enhance its functionality. It actually extends the MapReduce in order to include other capabilities like caching, task scheduler, loop awareness in scheduling tasks and so on. As many applications used by enterprises in the real wolrd need processing huge amount of data, Big Data in other words, it is essential to have data mining and extraction of business intelligence. For this purpose Hadoop is used as a distributed file system that supports MapReduce programming paradigm. This kind of programming paradigm is meant for processing big data in a distributed environment. Distributed programming frameworks like Hadoop, Haloop can help in achieving the task of processing Big Data in very less time. MapReduce is one of the scalable programming paradigm. The scalable feature of it is very important as data grows exponentially in the real word and that needs to be processed efficiently. Dryad is another MapReduce framework that is meant for processing Big Data. Many companies like Google, Facebook, Yahoo etc. are already using MapReduce programming paradigm in processing huge amount of data. The improved form of Hadoop is known as Haloop. Its architecture supports loop aware task scheduler and other features such as caching mechanism (Bu, et.al 2010).

Many data mining algorithms came into existence for mining data. These algorithms [2], [3], [5], [15], [16] are not able to handle uncertain data with huge datasets. In this paper we proposed a method that can handle user constraints and also uncertain data using MapReduce programming paradigm. Thus the proposed method can produce quality patterns that provide comprehensive business intelligence.

III. PROPOSED SYSTEM

The proposed system makes use of MapReduce programming model which runs in distriuted environment. The framework makes use of a distributed file system where thousands of nodes involve in processing big data. Moreover, the MapReduce programming framework can leverage the parallel power of modern processors thus producing high speed processing suitable for big data mining. Fiugure 2 shows how the overall flow takes place in the proposed system.

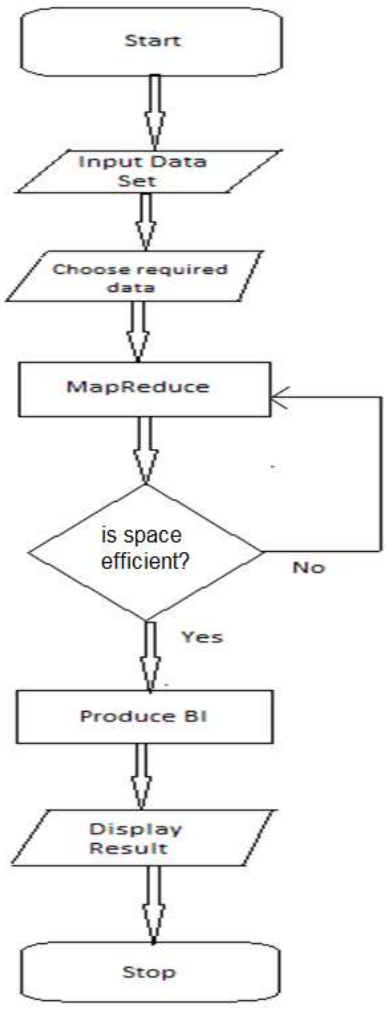

Figure 2 – Flow of the proposed system

As can be seen in Figure 2, it is evident that the proposed system makes use of MapReduce programming framework for big data mining. The propoed framework also takes user constraints besides processing big data using MapReduce programming model. The results contain comprehensive business intelligence which provides required knowhow to make well informed decisions. The underlying algorithm in the proposed system is outlined below.

 ISSN: 2321-8169
5090 - 5094

1 Input big data ***BD***
2 Input user constraints ***UC***
3 Pre-processing
4 Reduce search space iteratively based on ***UC***
5 Map task
6 Reduce task
7 Produce patterns
8 Return patterns

As can be seen in Figure 2, it is evident that the outlined algorithm takes care of the pre-processing and produce patterns that are based on the constraints. The results provided can give business intelligence that can be used to make well informed decisions.

IV. EXPERIMENTAL RESULTS

We built a prototype application that was used to perform experiments. The application demonstrates the usefulness of the proposed system. The exerimnets are made in terms of the time taken to process big data and other aspects like speedup, selectivity in the presence of user constraints.

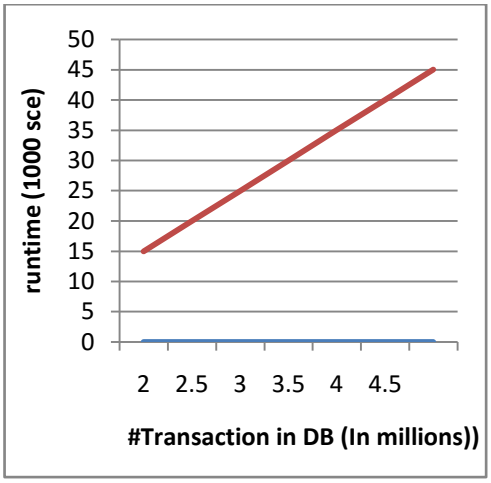

Fig 3 runtime vs. Transactions

As shown in figure 3 horizontal axis represents transaction in DB while vertical axis represents runtime.

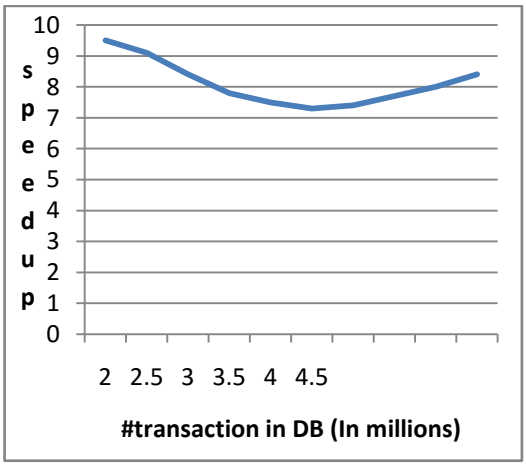

Fig 4 speedup vs. Transactions

As shown in figure 4 horizontal axis represents transaction in DB while vertical axis represents speedup.

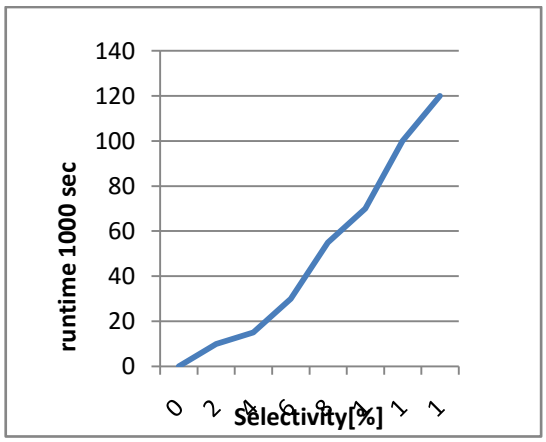

Fig 5 runtime vs. Selectivity.

As shown in figure 5 horizontal axis represents selectivity while vertical axis represents runtime.

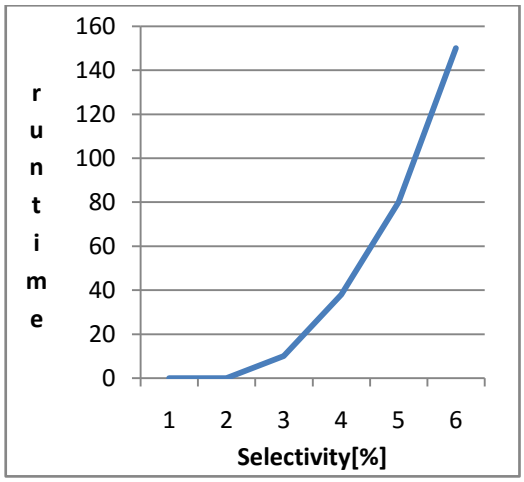

Fig 6 runtime vs. Selectivity

As shown in figure 6 horizontal axis represents selectivity while vertical axis represents runtime.

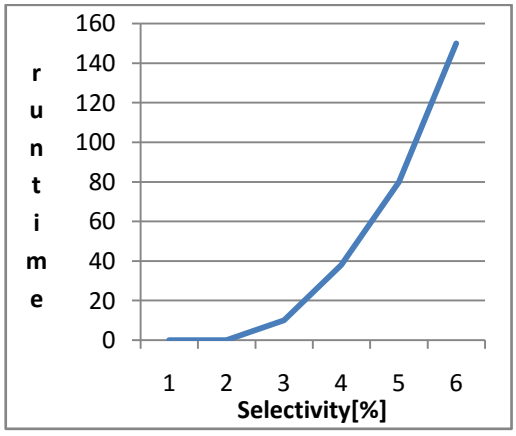

Fig 7 runtime vs. Selectivity

As shown in figure 7 horizontal axis represents selectivity while vertical axis represents runtime.

 ISSN: 2321-8169
5090 - 5094

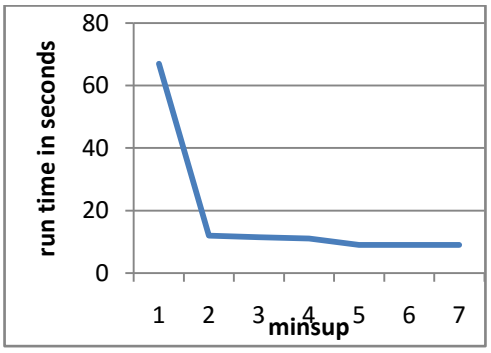

Fig 8 runtime vs. Minsup.

As shown in figure 8 horizontal axis represents minsup while vertical axis represents runtime.

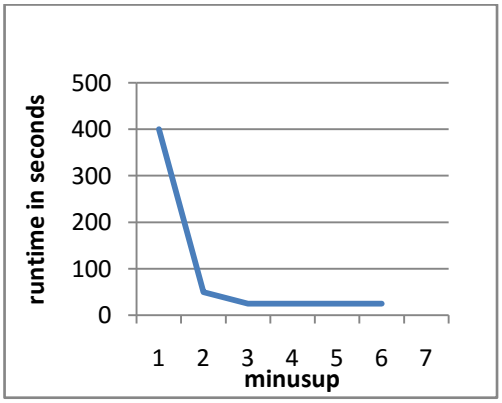

Fig 9 runtime vs. Minsup

As shown in figure 9 horizontal axis represents minsup while vertical axis represents runtime.

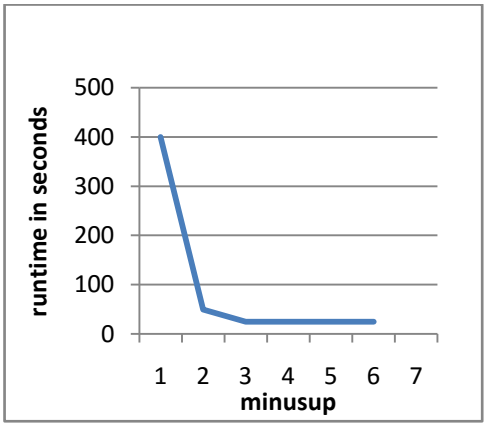

Fig 10 runtime vs. Minsup.

As shown in figure 10 horizontal axis represents minsup while vertical axis represents runtime.

V. CONCLUSION AND FUTURE WORK

This paper focuses on big data mining with reduced search space and deals with uncertain data as well. When search space is reduced, the processing efficiency of an algorithm can be improved. Towards this end, in this paper, we proposed a solution that works on uncertain data and reduces search space. Since the uncertain data implies the notion of "curse of dimensionality", the proposed system throws light on this for big data mining. MapReduce programming paradigm is used in order to handle huge amount of data. Since it is a distributed programming framework, it can leverage the parallel processing power of modern data centers and cloud computing. The proposed solution also has provision for users to select constraints with which quality of mining can be improved. We built a prototype application to demonstrate the proof of concept. The empirical results are encouraging. The proposed solution can be used to mine huge amount of data or big data with uncertain characteristics. This research can be extended further in order to improve the accuracy of producing patterns and secure the mapper from any sort of attacks.

REFERENCES

[1] P. Agarwal, G. Shroff, & P. Malhotra, "Approximate incremental bigdata harmonization," in IEEE Big Data Congress 2013, pp. 118–125.
[2] R. Agrawal & R. Srikant, "Fast algorithms for mining association rules," in VLDB 1994, pp. 487–499.
[3] A. Azzini & P. Ceravolo, "Consistent process mining over Big data triple stores," in IEEE Big Data Congress 2013, pp. 54–61.
[4] T. Condie, P. Mineiro, N. Polyzotis, & M. Weimer, "Machine learning for Big data," in ACM SIGMOD 2013, pp. 939–942.
[5] R.L.F. Cordeiro, C. Traina Jr., A.J.M. Traina, J. L ´opez, U. Kang, & C. Faloutsos, "Clustering very large multi-dimensional datasets with MapReduce," in ACM KDD 2011, pp. 690–698.
[6] J. Dean & S. Ghemawat, "MapReduce: simplified data processing on large clusters," CACM 51(1): 107–113, Jan. 2008.
[7] A. Koufakou, J. Secretan, J. Reeder, K. Cardona, & M. Georgiopoulos, "Fast parallel outlier detection for categorical datasets using MapReduce," in IEEE IJCNN 2008, pp. 3298–3304.
[8] A. Kumar, F. Niu, & C. R´e, "Hazy: making it easier to build and maintain Big-data analytics," CACM 56(3): 40–49, Mar. 2013.
[9] L.V.S. Lakshmanan, C.K.-S. Leung, & R.T. Ng, "Efficient dynamic mining of constrained frequent sets," ACM TODS 28(4): 337–389, Dec. 2003.
[10] C.K.-S. Leung, "Frequent itemset mining with constraints," in Encyclopedia of Database Systems, pp. 1179–1183, 2009.
[11] C.K.-S. Leung, "Mining uncertain data," WIREs Data Mining and Knowledge Discovery 1(4): 316–329, July/Aug. 2011.
[12] C.K.-S. Leung & F. Jiang, "Frequent itemset mining of uncertain data streams using the damped window model," in ACM SAC 2011, pp. 950– 955.

ISSN: 2321-8169
5090 - 5094

[13] C.K.-S. Leung & F. Jiang, "Frequent pattern mining from time-fading streams of uncertain data," in DaWaK 2011 (LNCS 6862), pp. 252–264.

[14] C.K.-S. Leung, M.A.F. Mateo, & D.A. Brajczuk, "A tree-based approach for frequent pattern mining from uncertain data," in PAKDD 2008 (LNAI 5012), pp. 653–661.

[15] C.K.-S. Leung & S.K. Tanbeer, "Fast tree-based mining of frequent itemsets from uncertain data," in DASFAA 2012 (LNCS 7238), pp. 272– 287.

[16] C.K.-S. Leung & S.K. Tanbeer, "PUF-tree: A compact tree structure for frequent pattern mining of uncertain data," in PAKDD 2013 (LNCS 7818), pp. 13–25.