The Research of Distributed Data Mining Knowledge Discovery Based on Extension Sets

Vuda Sreenivasarao
Assoc Professor & Head
CSE & IT Department
St.Mary's, Hyderabad
Andhra Pradesh, India

Rallabandi Srinivasu
Professor& Director
St. Mary's Group of Institutions
Hyderabad
Andhra Pradesh, India

Prof. G.Ramaswamy
Professor & Director
Priyadarsini Engineering College
Tenali
Andhra Pradesh, India

Nagamalleswara Rao Dasari
Assistant Professor, CSIT Department
St. Mary's College of Egg. & Tech
Andhra Pradesh, India

Dr. S Vidyavathi
Associate Professor, CSIT Department
JNT University, Hyderabad
Andhra Pradesh, India

ABSTRACT

Distributed Data Mining(DDM) has evolved into an important and active area of research because of theoretical challenges and practical applications associated with the problem of extracting, interesting and previously unknown knowledge from very large real-world databases. Extension Set Theory is a mathematical formalism for representing uncertainty that can be considered an extension of the classical set theory. It has been used in many different research areas, including those related to inductive machine learning and reduction of knowledge in Distributed data-based systems. Extenics is a theory to solve the contradiction problem, it will be a new way to look for and find knowledge through analysis the contradiction and transformation the result of the data mining using the extension methods. In this paper, introduced the matter-element and extension set that is the base of the extenics, researched the way to find out and generate the new knowledge that help by the divergence, change and transformation based on the extension The main aim is to show how Extension sets can be effectively used to extract knowledge from large databases.

Keywords:—Data mining, Data tables, Distributed Data Mining (DDM), Extension sets.

1. INTRODUCTION:

Data mining technology has emerged as a means for identifying patterns and trends from large quantities of data. Data mining is a computational intelligence discipline that contributes tools for data analysis, discovery of new knowledge, and autonomous decision making. The task of processing large volume of data has accelerated the interest in this field. As mentioned in Mosley (2005) data mining is the analysis of observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

Data mining has gained popularity in the database field recently; it has been mostly used by statisticians, data analysts and so on. Data mining techniques can be divided into five classes of methods: predictive modeling; clustering; data summarization, change and deviation detection. Some of these techniques are beginning to be scaled to operate on databases.

Distributed Data Mining (DDM) aims at extraction useful pattern from distributed heterogeneous data bases in order, for example, to compose them within a distributed knowledge base and use for the purposes of decision making. A lot of modern applications fall into the category of systems that need DDM supporting distributed decision making. Applications can be of different natures and from different scopes, for example, data and information fusion for situational awareness; scientific data mining in order to compose the results of diverse experiments and design a model of a phenomena, intrusion detection, analysis, prognosis and handling of natural and man-caused disaster to prevent their catastrophic development, Web mining ,etc. From practical point of view, DDM is of great concern and ultimate urgency.

Extension is a theory invented by Cai Wen, a Chinese scholar, to resolve the conflict questions. it study that the disciplinarian and the ways to expand things and innovation, it is across the natural and social sciences. Extension think that everything in the world can be expanded, the character that the tings can be expanded they call it is extension. By using symbols to indicate the extension of things, you can use computers to help people deal with the contradictions. and refer to the problem-solving strategies, ideas, methods, and so on. This theory and methods combine with computer, can greatly expand the scope of application of computers, and also can expand human wisdom

and capacity greatly. The method of Extension can be used in the field of knowledge discovery to analyze and transform conflicts the result of data mining in order to discover new knowledge, the way that using extension to discover knowledge applied to the decision-making areas, will be find new strategies, applied to product manufacturing, will be new product design, applied to commercial areas, will be new processes. and so on.

Knowledge discovery in databases (data mining) is one of the most promising directions spanning database and artificial intelligence research. Data mining is a nontrivial extraction of potentially significant facts from data. The methodology of extension sets offers some framework to study data mining problems. Using the extension set approach to discovery of strong rules in stock market data is very interesting as indicated in papers. It is necessary to make some discrimination of continuous attributes, before generation of rules.

The proposed approach:

- Provides efficient algorithms for finding hidden patterns in data,
- Finds minimal sets of data (data reduction),
- Evaluates significance of data,
- Generates sets of decision rules from data,
- It is easy to understand,
- Offers straightforward interpretation of obtained results,

The remaining sections of the paper are organized as follows. In Section II we describe Architecture of proposed system .In Section III we describe Distributed Data Mining. In Section IV we describe matter-element and extension sets. In Section III we describe properties and In Section VI concludes the paper.

2. ARCHITECTURE OF PROPOSED SYSTEM:

The overview of the architecture of the system can be seen in figure. The proposed architecture will adopt the traditional architecture of a data mining system. Data from multiple channels is collected on the operational data store for fast transaction and up to date data that can be used for the front office. Then, periodically, the data is extracted, cleans, transformed and imported into the data warehouse. The data will then will be send to the appropriate data marts for departmental use. Then, according to the needs of the user, either the enterprise data or the departmental data is sent to the OLAP tier for processing. The results is then stored and then sent to the decision makers through the use of thin clients. The overview of this architecture is seen in Figure 1. The proposed system is pretty good in theory as it provides compartmentalization of data and collection of data from multiple channels. The architecture is simple and sticks to the basis of founded work and should provide a good base for the system.

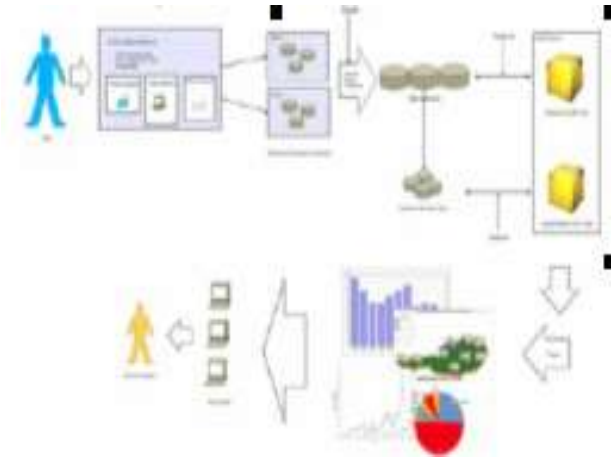

fig 1: The over view of proposed system Architecture

3. DISTRIBUTED DATA MINING:

Data mining technology has emerged as a means for identifying patterns and trends from large quantities of data. Data mining and data warehousing go hand-in-hand: most tools operate on a principal of gathering all data into a central site, then running an algorithm against that data (Figure 2). There are a number of applications that are infeasible under such a methodology, leading to a need for distributed data mining.

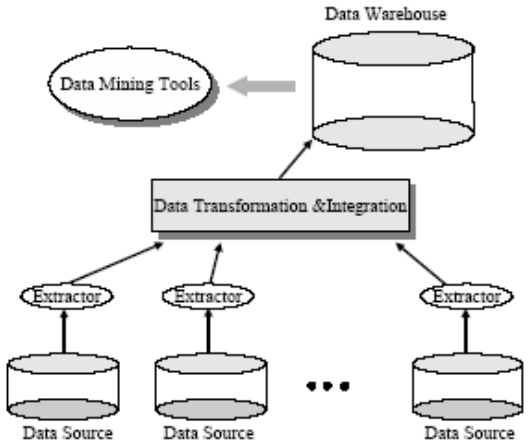

Fig 2. A Data Warehouse Architecture.

Distributed data mining (DDM) considers data mining in this broader context. As shown in figure(3), objective of DDM is to perform the data mining operations based on the type and availability of the distributed resources. It may choose to download the data sets to a single site and perform the data mining operations at a central location.

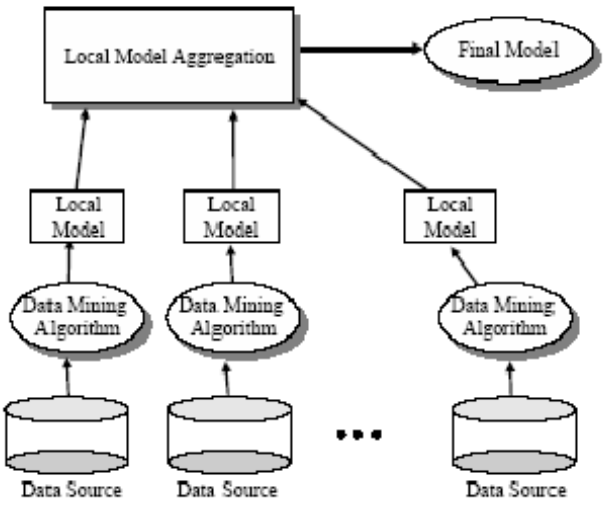

Fig 3. A Distributed Data mining Frame work.

Data mining is a powerful new technology with great potential to help companies focus on the most important information in the data they have collected about the behavior of their customers and potential customers. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods. It discovers information within the data that queries and reports can't effectively reveal.

4. MATTER-ELEMENT AND EXTENSION SETS:

4.1. Definition 1: Matter-element：

Triple set R = (N, c, v) that element in it is order to describe things as the basic element, referred to as

matter- element. Among them, N means things, c refer to the characteristics, v describe the value of N on the characteristics of c . These elements as the three elements of matter-element. The concept of matter-element, v=c(N) reflects the relationship between the quality and quantity of things. Put forward the concept of characteristic element M(c, v) , which include the name of character c and the corresponding value v .it described the character that is often said by human. A object can have many characteristics element, it can described by Many dimension matter-element .if have dynamic matter-element, the formula R(t)=(N(t),c, v(t))=(N(t),c, c(N(t))) described that the things change follow the time.

As the define of matter-element, it also exist the event-element and relationship-element. Described it with I = (d, b, u) and Q= (s, a, w) . matter-element, event-element and relationship-element in harness called by base-element use their and complex-element that assembled by their can describe the kinds of question in our world using the formalization way. in order to use the simple way to describe the question. we only use the matter-element in this paper.

4.2 . Definition 2: the extension of matter-element:

Matter-element with a divergent, conjugate, relevance, including and extending, this five kinds of Characterize called the extension of matter-element. The extension of Matter-elementc can make the thing extend to the more possible from the different directions, such as analysis thing to outside, to inside, parallel , flexible decomposition and composition .based on these methods, we can innovate and to solve the conflict problem.

4.3. Definition 3: matter-element transformation：

Matter-element R0 =(N0,c0,v0) to change the matter-element R =(N, c, v) or a number of matter-element

R 1 =(N1,c1,v1), R 2=(N2,c2,v2),------- R n =(Nn, cn, vn), called the matter-element transformation, matter-element transformation with four basic types, that is the replacement, decomposition, additions and deletions, and it has four basic operations, that is product, contradictorily .or and.

4.4. Definition 4: Extension Set：

If U is domains, k is the mapping of U to real domain I , T is a transformation of a given element of the U , we call (T)={(u,y,y')|uϵU,y=k(u)ϵI,y'=k(Tu)ϵI} that is a extension set which is about T transformation on the domain U . y=k(u) is the associated function e of (T) In order to directly describe the nature of things is exchange, and the process of qualitative change and quantitative change, extension set use the number between (−,+) to describe the degree of a certain nature of things ,and to describe the transformation of things between "is" and "no" get across the extension. based on the associated function and extension transformation of the state of the things characteristics that include the transformation of domains, transformation of associated function and transformation of element, we can divide the domains into three parts: the stability domain, the extension domain and extension sector. While T=e(e is identical transformation),it called static extension set. transform element extension set (static extension set) can be divided into three parts sector:

positive domain, negative domain and zero-boundary. While T≠ e, it called dynamic extension set, transform element extension set can be divided into five parts: positive extension domains, negative extension domain, positive stable domains, negative stable domain and extension boundary.

Defining the name of a matter by O, one of the characteristics of the matter by c and the value of c by v , a matter-element in extension theory can be described as follows:

M= (O, c, v) Where O, c and v are called the three fundamental elements of the matter-element. For example, M = (Tang, Weight, 60 kg) can be used to state that Tang's weight is 60 kg. If the value of the characteristic has a classical domain or a range, we define the matter-element for the classical domain as follows:

M= (O, c, v) = (o, c, ‹vl , vll›

Where vl and vll are the lower bound and upper bound of a classical domain. Relevant rule is defined that certain cases can bring about others cases. Such as rule X → Y , X and Y are the attribute variables in database.

Extension relevant rule with matter-element is

$$\bigwedge_{i=1}^{n} r_i \Rightarrow (l)R$$

Relevant rules with combined type are rules which have essence-element item and extension transform item. It is r1∩ r2∩---------- ∩ r n =(l)R

Relevant rule with combined type is fit for researching relevant rule of complicated system.

4.5. Extension transformation:

Extension transformation is the tool for solving the problem. It can turn the unknowable problem into knowable problem, turn the unfeasible problem in feasible problem, turn the false proposition into true proposition and turn the incompatible problem into compatible problem. Extension transformation is turning one object into another one, it is turning base element u into base element v , and it can be represented as: Tu = v

Extension information is the information that is used to solve the incompatible problem. Base element in extenisc, including matter-element, case-element and relation element, is the basic information of extension information. Transformation in extenisc is transforming the information, so that transforming the incompatible problem into compatible problem.

Extension information = base element + extension transformation.

The basic information in extension information is static description, but transformation of the information has the character of variety. It must be using the extension transformation and variable information to solve the incompatible problem.

4.6. The concept of extension distributed data mining:

Extension data mining that mining extension knowledge is the extension of data mining. Data mining can acquire knowledge (condition result), extension transforming the condition and conducting transforming the result can acquire variable knowledge (extension knowledge): T condition → T result It calls this variable knowledge as extension data mining.

4.7. The Divergence of distributed data mining Knowledge Generation:

The thing can have a variety characteristic, a characteristic also can be haven by a number of things, so, while we have the condition matter-element and the goal matter-element, we can divergence and inference matter-element get it numbers characteristic, then get numbers things helped by one characteristic, by this way, realization the knowledge diverge

R┤{R1 │ R1 =(N, ci, v),I = 1,2,3,……,n,c1€E(c)} (1)

Formula (1) display that there are many characteristics of the thing, referred to as "one thing more characteristic", c means characteristics, it can provide a number of characteristics C1, C2, C3,…… .. C n. symbols ┤ means inference, E(c) of all of the characteristics.

R┤{R1 │ R1 =(Ni, c, v1),I = 1,2,3,……,n.} (2)

Formula (2) display that a characteristic can be had by a number of things that have the different eigen value, referred to as "one characteristic, many things, many values",

R┤{R1 │ R1 =(Ni, ci, v),I = 1,2,3,……,n.} (3)

Formula (3) display a value that can be get from many characteristics of many things. referred to as "one

value, many characteristics, many things,"

R┤{R1 │ R1 =(Ni, c, v),I = 1,2,3,……,n.} (4)

Formula 4 display that different things can have the same characteristics and value. referred to as "one

characteristic, many things," Conjunction the formula (1)(2)(3),the result is the divergence of knowledge(%):

R┤{ (N, ci, v),I = 1,2,3,……p,} U { (Ni, ci, v),I = 1,2,3,……q,} U{ (Ni, c, vi),I = 1,2,3,……m,}= W1(R) U W2(R) U W3(R) (5)

4.8. Selection Process of the distributed data mining Knowledge Discovery Based on Extension sets:

The method of the knowledge discovery based on the extension is a method that based on the extension theory, simulated human thought patterns, using formal, quantitative way to solve the incompatibility problem to find out the new knowledge. the basic process is:

a) define the objectives and conditions of the practical problems.

b) set up the extension model using mater-element.

c) establish detection function for the paradoxical problem: compatibility function. using the method as build the associated function as to build the compatibility function K(p) according to the requirement to satisfy realization the goal under the condition of the practical problems. if K(p) ,call the problem P is incompatible problem. if K(p), call the problem P is critical problem. to the critical problem, we can inference using by the extension or other mathematics methods.

d) to the incompatible problem, need to transfer incompatible into compatible helped by the extension methods that including divergence, change and transformation. at finally generate new knowledge.

e) through the evaluation of the optimal to evaluate and select knowledge that found by extension methods.

5. PROPERTIES:

1. The space H and each Tn-set is an extension set of order n.

2. If [M] is an arbitrary collection of sets Jn then their intersection is also a Jn.

3 If M is a Jn and X is a Cn then M-X is a Cn

4 For any set A subset of Ü the set Δn(A) is a Jn .

5 If the subspace X is not cut by any Tn then Δn(X) also has this property.

6. Each set Bn is also a set Jn

7 If H is a space of type V then each Jn is a Jn+i.

8 Let M be a Jn and letf:X→*Sn where X is a closed subset of M. If N is an essential membrane for f then N subset of.M.

9 . If M is a Jn, X a subset of M and N a Cn irreducible about X, then N subset of M.

10. For any property P the property of being a P-endelement is inductive.

11. Any Tn-endelement of H is a Jn.

12. If H is a Cn then any Tn- endelement is a Cn.

5.1. Further properties of extension sets:

1 .For any subset Z subset of H we haveΔn(A) =Δn(Ā).

2 .If M is a Jn, Za Tn and M - Z = M1 + M2 is a separation then Mi+Z is a Jn.

3 .If x and y are points of H then neither x nor y is a cut point of Δn(x+y) if this set is a continuum.

4 .Let N be a subspace the complement of which is the union of a collection of pair-wise disjoint open sets whose boundaries are sets Tn. Then N is a Jn.

5. If X is a subspace then each Tn -endpoint of Δ(X) is a Tn-endpoint of X.

6 .If M+N and M . N are Jn-sets then so also are M and N if they are sub spaces.

7 .If X is a Co then Δ(X) is a Co.

8. Let Z be the set of all points of H that are not Tn-endpoints. Then, if H is a Cn, the set Z is n-connected.

6. CONCLUSION :

Distributed data mining knowledge discovery Based on the extension set is a way to solve the incompatibility problem .Describe the problem and the goal using matter-element, and then divergence, change and transformation these matter-element make them come into compatibility. Through the evaluate that using mathematics methods to find the optimal knowledge in the object. In this paper, researched the methods of the divergence and transformation that based on the extension sets, build the model to find the optimal knowledge, design procession of the knowledge discovery. The Inadequacies is that it is too many matter-element need to list to give a application. So, This article only provides a methods to realization distributed data mining knowledge discovery based on extension stes.

7.REFERENCES:

[1] Wen Cai,"The Extension Set And Non-Compatible Problem",science exploration,1983,3(1):83

[2] Wen Cai ,"Matter-Element Models And Their Application",science and technoloty documentation publishers,bejing 1994.

[3] Yinai Sun,Yongquan Yu,Wen Cai And Guangqiang Li."Extension Set And The Dependent Function Of System Model",ICICIC'06,0-7695-2616-0/06.

[4] Xiaoyuan Zhu,Yongquan Yu,Hong Wang,Yunhua Chen."Extension Set And The Application In Datamining".ICCIT,2007.28 220

[5]. Cai Wen. Extension theory and its application, Chinese Science Bulletin. 44(17), 1999:1538-1548

[6]. Chen Wenwei.Research on mining the mutative knowledge with extension data mining. Engineering Science, 8(11), 2006:70-73

[7]. WenWei Chen, JinCai Huang .Extension Transformation and Extension Knowledge Representation of Attribute Reduction and Date Mining .Journal of Chongqing Institute of Technology ,2007:1-4

[8]. Li Lixi, Yang Chunyan, Li Huawen, Extension Strategy Generating System, Science Press, Beijing, 2006.

[9]. J.Han and M.Kamber. "Data mining: concepts and techniques", Morgan Kaufman Publishers, San

Francisco, CA. 2001.

[10]. Pawlak Z. Why rough sets: proc.of the 5th IEEE International Conference on Fuzzy Systems[C]. New Orleans:IEEE,1996:738□743.

[11] Vuda sreenivasarao ,Dr S.Vidyavathi : Distributed data mining and mining multi – agent data base :IJCSE:2010 : 1237-1244.

[12] B. Lloyd. Been gazumped by Google? Trying to make sense of the \Florida" update. Search Engine Guide,

November 25, 2003.

[13] M. V. Mahoney and P. K. Chan. Learning Non stationary models of normal network trac for detecting novel attacks. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 376{385, Edmonton, Canada, 2002. ACM Press.

[14] P. Robertson and J. M. Brady. Adaptive image analysis for aerial surveillance. IEEE Intelligent Systems, 14(3):30{36, 1999.

[15] T. Senator. Ongoing management and application of

Discovered knowledge in a large regulatory organization: A case study of the use and impact of NASD regulation's advanced detection system (ADS). In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 44{53, Boston, MA, 2000. ACM Press.

[16] A. Hurson, M. Bright, and S. Pakzad, Multidatabase systems: an advanced solution for global information sharing. IEEE Computer Society Press, 1994.

[17] H. Liu, H. Lu, and J. Yao, Identifying Relevant Databases for Multidatabase Mining. In: Proceedings of Pacific-Asia

Conference on Knowledge Discovery and Data Mining, 1998: 210–221.

[18] J. Yao and H. Liu, Searching Multiple Databases for Interesting Complexes. In: Proc. of PAKDD, 1997: 198-210.

[19] X.Wu and S. Zhang, Synthesizing High-Frequency Rules from Different Data Sources, IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 2, March/April 2003: 353-367.

[20] C. Zhang and S. Zhang, Association Rules Mining: Models and Algorithms. Springer- Verlag Publishers in Lecture Notes on Computer Science, Volume 2307, p. 243, 2002.

[21] N. Zhong, Y. Yao, and S. Ohsuga, Peculiarity oriented multi-database mining. In: Proceedings of PKDD, 1999: 136-146.

[22]. Date, C. An Introduction to Database Systems, Volume I, The Systems Programming Series, Addison-Wesley, 1986.

[23]. Ullman, I.D. Principles of Databases and Knowledge-Based Systems, Volume I, Computer Science Press, 1988

[24]. Ullman, J., Widom, J., A First Course in Database Systems, Prentice Hall, 2001

ABOUT THE AUTHORS

Vuda Sreenivasarao received his M.Tech degree in Computer Science & Engg from the Satyabama University, in 2007.Currently working as Associa Professor & Head in the Department of Information Technology(IT) at St.Mary's college of Engineering & Technology, Hyderabad, India. He is Currently Pursuing the PhD degree in CSIT Department at JNT University, Hyderabad, India. His main research interests are Data Mining, Network Security, and Artificial Intelligence. He has got 10years of teaching experience .He has published 14 research papers in various international journals. He is a life member of various professional societies like MIACSIT, MISTE and MIAENG.

Rallabandi Srinivasu Received his M.Sc Degree from Nagarjuna University Campus in 2000, M.Phil degree from Acharya Nagarjuna University, Guntur .in 2009. He is currently Pursuing Ph.D in Management from Rayalaseema University, India. Currently working as Director of St.Mary's PG College, Hyderabad, India. His main research interests are Data Mining, Management Information Systems.

Prof.G.Ramaswamy received the M.Tech degree in Information Technology from the Punjabi University, in 2003. He received the Ph.D degree in Computer Science & Engg from Magadh University from the 2007. Currently he is professor and director in Priyadarsini Engineering College. His research interests include Network Security, Cryptography.

Nagamalleswara Rao Dasari received B.Tech., degree in Computer Science and Information Technology from JNTU in 2006. He is a research scholar in CSIT department, St.Mary's Engg. College, Hyderabad, Andhra Pradesh, India. His research interests include Network Security, Cryptography, Data Mining. He is a member of IAENG.

Dr. S Vidyavathi received her PhD degree from IIT Mumbai, She is currently working as Associate Professor in CSIT Department , JNT University, Andhra Pradesh, India.