A Comprehensive Analysis on Recommendation System for E-Commerce based on Distributed Data Mining

Bhanu Priya G, Nibedita P
Dept of CS&E,Dept of CS&E,
Vivekananda Institute of Technology,Vivekananda Institute of Technology
Bangalore-560074,IndiaBangalore-560074,India

***Abstract*-Data mining technology has emerged as a means for identifying patterns and trends from large quantities of data in many fields [such as research, computer science, E-commerce etc...] The explosive growth of e-commerce and online environments has made the issue of information search and selection increasingly serious, users are overloaded by many options to consider and they may not have the time or knowledge to personally evaluate these options. The field of Distributed Data Mining (DDM) deals with these challenges in analyzing distributed data. In this paper we survey the association rule mining along with the recommendation system to recommend the products for the end users.**

Keywords: Distributed data mining, recommendation system, association rule mining, a priori algorithm

I. INTRODUCTION

The continuous developments in information and communication technology have recently led to the appearanceof distributed computing environments, which comprise several, and different sources of large volumes of data and severalcomputing units. The most prominent example of a distributed environment is the Internet, where increasingly moredatabases and data streams appear that deal with several areas, such as meteorology, oceanography, economy and others.

In a world where the number of choices can be overwhelming, the recommendersystems help users to find and evaluate items of interest. They connect users withitems to “consume” (purchase, view, listen to, etc.) by associating the content ofrecommended items or the opinions of other individuals with the consuming user’sactions or opinions. Such systems have become powerful tools in domains fromelectronic commerce to digital libraries and knowledge management. For example,a consumer of just about any major online retailer who expresses an interest in anitem – either through viewing a product description or by placing the item in his“shopping cart” – will likely receive recommendations for additional products.These products can be recommended based on the top overall sellers on a site, onthe demographics of the consumer, or on an analysis of the past buying behavior ofthe consumer as a prediction for future buying behavior.

In the distributed data mining(DDM)[35,36] to recommend the products or in E-commerce the Association rules have been used for many years . An association ruleexpresses the relationship that one product is often purchased along with otherproducts. The number of possible association rules grows exponentially with thenumber of products in a rule, but constraints on confidence and support, combinedwith algorithms that build association rules[3] with itemsets of n items from rules withn-1 item itemsets, reduce the effective search space. Association rules can form avery compact representation of preference data that may improve efficiency ofstorage as well as performance. They are more commonly used for largerpopulations rather than for individual consumers, and they, like other learningmethods that first build and then apply models, are less suitable for applicationswhere knowledge of preferences changes rapidly. Association rules have beenparticularly successfully in broad applications such as shelf layout in retail stores.By contrast, recommender systems[2,3,] based on collaborative filtering(CF)[5,6] techniques are easier to implementfor personal recommendation in a domain where consumer opinions are frequentlyadded, such as on-line retail.

II. RELATED WORK

The development of the Internet, the problem of information overload is becoming increasing serious. People all have experienced the feeling of being over whelmed by the large number of data. Many researchers pay more attention on building a proper tool which can help users obtain personalized resources. Recommendation systems are one such software tool used to help users obtain recommendations for unseen items based on their preferences.

By Guangping Zhuo, Jingyu Sun and Xueli Yu[10]“A Framework for Multi-Type Recommendations” deals in the field of web mining concern on some drawbacks in collaborative filtering and also on multi type

recommendation. Collaborative filtering (CF) is an effective method of recommender systems (RS) has been widely used in online stores. However, CF suffers some weaknesses: problems with new users (cold start), data sparseness, and difficulty in spotting "malicious" or "unreliable" users and so on. Additionally CF can't recommend different type items at the same time. So in order to make it adaptive new Web applications, such as urban computing, visit schedule planning and so on, introduced a new recommendation framework, which combines CF and case-based reasoning (CBR) to improve performance of RS. Based on this framework, the authors have developed a semantic search demo system—our Visit, which shows that proposed framework is an effective recommendation model. Two key algorithms, MIFA and RAA, are used. Additionally, authors have validated them using an application instance, which is a demo system for recommending multi type recommendations combining CF and CBR. Advantage of this method is that it involves a few of cases in the online and adjusts the rating of main items through associative other type items in order to find fit recommendations.

By Yechun Jiang, Jianxun Liu, Mingdong Tang and Yechun Jiang, Jianxun Liu, Mingdong Tang [2] "An Effective Web Service Recommendation Method based on Personalized Collaborative Filtering", Describing an effective personalized collaborative filtering method for Web service recommendation. A key component of Web service recommendation techniques is computation of similarity measurement of Web services. Different from the Pearson Correlation Coefficient (PCC) similarity measurement, they take into account the personalized influence of services when computing similarity measurement between users and personalized influence of services. Based on the similarity measurement model of Web services, develop an effective Personalized Hybrid Collaborative Filtering (PHCF) technique by integrating personalized user-based algorithm and personalized item-based algorithm. Also conduct series of experiments based on real Web service QoS dataset WSRec which contains more than 1.5 millions test results of 150 service users in different countries on 100 publicly available Web services located all over the world. Experimental results show that the method improves accuracy of recommendation of Web services significantly.

By Zan huang,Hsinchun chen,and Denial zeng[3] "Applying associative retrival technique to alleviate the sparsity problem in collaborative filtering" Recommender systems are being widely applied in many application settings to suggest products, services, and information items to potential consumers. Collaborative filtering, the most successful recommendation approach, makes recommendations based on past transactions and feedback from consumers sharing similar interests. A major problem limiting the usefulness of collaborative filtering is the sparsity problem, which refers to a situation in which transactional or feedback data is sparse and insufficient to identify similarities in consumer interests. In this article, we propose to deal with this sparsity problem by applying an associative retrieval framework and related spreading activation algorithms to explore transitive associations among consumers through their past transactions and feedback. Such transitive associations are a valuable source of information to help infer consumer interests and can be explored to deal with the sparsity problem.

By Luo Zhenghua, [1]"Realization Of Individualized Recommendation System On Books Sale" In order to recommend potential books of interest to customers efficiently, the association rules in data mining to e-commerce business systems of book sales, designs an individualized recommendation system of book sales, and introduces the flow of the recommendation system and the specific realization procedures of data input, data preprocessing, association rules existence and individualized recommendation. Results show that the web site based on this article has shown great performance.

By Agrawal, R., Imielinski, T., Swami, A. N.[18] "Mining association rules between sets of items in large databases" The system aims at developing pattern mining and prediction techniques that explore the correlation between the moving behaviors and purchasing transactions of mobile users to explore potential M-Commerce features.This enables the businesses to understand the patterns hidden inside past purchase transactions, thus helping in plan and launch new marketing campaigns in prompt and cost effective way. The following illustrates several applications in sale and marketing.

Application is used for market basket analysis to provide insight information on what product combinations were purchased, when they were bought and in what sequence by customers. This information helps businesses to promote their most profitable products to maximize the profit. In addition, it encourages customers to purchase related products that they may have been missed or overlooked. Retails companies can make use of this application to identify customer's behavior buying patterns.

Other technologies have also been applied to recommender systems, including Bayesian networks, clustering, and Horting. Bayesian networks create a model based on a training set with a decision tree at each node and edges representing user information. The model can be built offline over a matter of hours or days. The resulting model is very small ,very fast, and essentially as accurate as nearest neighbor methods. Bayesian networks may prove practical for environments in which knowledge of user preferences changes slowly with respect to the time needed to build the model but are not suitable for environment in which user preference models must be updated rapidly or frequently.

Clustering techniques work by identifying group of users who appear to have similar preferences. Once the clusters are created, predictions for an individual can be made by averaging the opinions of the other users in that cluster. some clustering techniques represent each user with partial

participation in several clusters. The prediction is then an average across the clusters, weighted by degree of participation clustering technique usually produce less personal recommendation than other methods , and in some cases, the cluster have worse accuracy than nearest neighbor algorithms. Once the clustering is complete ,however performance can be very good ,since the size of the group that must be analyzed is much smaller .clustering technique can also be applied as a first step for shrinking the candidate set in a nearest neighbor algorithm or for distributing nearest neighbor computation across several recommender engines while dividing the population into clusters may hurt the accuracy or recommendation to users near the fringes of their assigned cluster, pre clustering may be a worthwhile tradeoff between accuracy and throughput .
To overcome all these problems in a web pages the stronger tool called recommendation system based on association rule mining in distributed data mining has been invented by many researchers.

DDM (distributed data mining)is a complex system focusing on the distribution of resources over the network as well as data mining processes. The very core of DDM systems is the scalability as the system configuration may be altered time to time,
Therefore designing DDM systems deals with great details of software engineer issues, such reusability, extensibility ,and robustness. For these reasons, agents' characteristics are desirable for DDM systems.Furthermore, the decentralization
property seems to fit best with the DDM
requirement. At each data site, mining strategy is deployed specifically for the certain domain of data.

III. DISTRIBUTED DATA MINING

Data mining deals with the problem of analyzing data in scalablemanner. DDM is a branch of the field of data mining that offers a framework tomine distributed data paying careful attention to the distributed data and computingresources..

The development of data mining algorithms that work well under the constraintsimposed by distributed datasets has received significant attention from the data miningcommunity in recent years. The field of DDM [35,36]has emerged as an active area of study.Some features of a distributed scenario where DDM is applicable are as follows.
1. The system consist of multiple independent sites of data and computation which
Communicate only through message passing.
2. Communication between the sites is expensive.
3. Sites have resource constraints *e.g.* battery power.
4. Sites have privacy concerns.

DDMalgorithms (association rule mining, clustering ,classification,preprocessing, *etc.*), systems issues in DDM (security, architecture, *etc.*), and in parallel data mining. Many of the DDM applications deal with continuousdata streams. Therefore, developing DDM algorithms that can handle such stream scenariosis becoming increasingly important.

Traditional warehouse-based architectures for data mining suppose to have centralized data repository. Such a centralized approach is fundamentally inappropriate for most of the distributed and ubiquitous data mining applications. In fact, the long response time, lack of proper use of distributed resource, and the Fundamental characteristic of centralized data mining algorithms do not work well in distributed environments. A scalable solution for distributed applications calls for distributed processing of data, controlled by the available resources and human factors.

Figure 1 a general Distributed Data Mining framework is presented. In essence, the success of DDM algorithms lies in the aggregation. Each local model represents locally coherent patterns, but lacks details that may be required to induce globally meaningful knowledge.
For this reason, many DDM algorithms require a centralization of a subset of local data to compensate it. The ensemble approach has been applied in various domains to increase the accuracy of the predictive model to be learnt. It produces multiple models and combines them to enhance accuracy. Typically, voting (weighted or un-weighted) schema are employed to aggregate base model for obtaining a global model.

IV. ASSOCIATION RULE MINING

An association rule is a rule which implies certain association relationships among a set of objects (such as ``occur together'' or ``one implies the other'') in a database. Given a set of transactions, where each transaction is a set of literals (called items), an association rule[3,33,34]is an expression of the form X Y , where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y .

Apriori Algorithm
An association rule mining algorithm, Apriori has been developed for rule mining in large transaction databases by IBM's Quest project team . An *itemset* is a non-empty set of items.
They have decomposed the problem of mining association rules into two parts

Find all combinations of items that have transaction support above minimum support. Call those combinations frequent item sets.

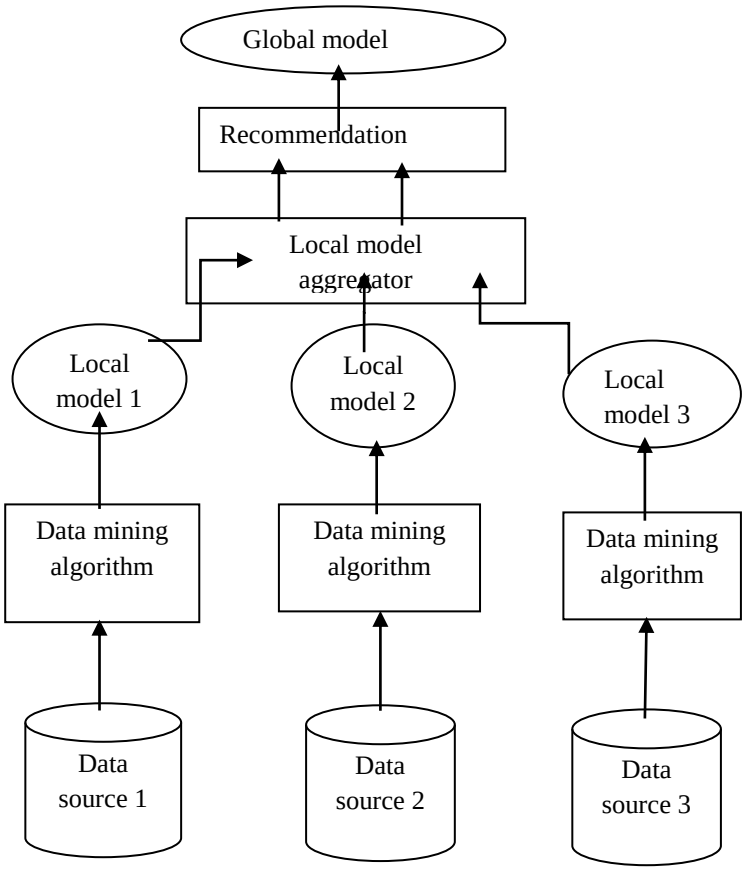

Fig 1: Distributed data recommendation procedure

Use the frequent itemsets to generate the desired rules. The general idea is that if, say, ABCD and ABare frequent itemsets, then we can determine if the rule AB CD holds by computing the ratio r= support(ABCD)/support(AB). The rule holds only if r >= minimum confidence. Note that therule will have minimum support because ABCD is frequent. The algorithm is highly scalable .The Apriori algorithm used in Quest for finding all frequent itemsets is given below.

procedure AprioriAlg()

begin
L1 := {frequent 1-itemsets};
for (k := 2; Lk-1 0; k++) **do** {
Ck= apriori-gen(Lk-1) ; // new candidates
for all transactions t in the dataset **do** {
for all candidates c Ck contained in t **do**
c:count++
}
Lk = { c Ck | c:count >= min-support}
}
Answer := k Lk
End

It makes multiple passes over the database. In the first pass, the algorithm simply counts item occurrencesto determine the frequent 1-itemsets (itemsets with 1 item). A subsequent pass, say pass k, consists of twophases. First, the frequent itemsets Lk-1 (the set of all frequent (k-1)-itemsets) found in the (k-1)th pass areused to generate the candidate itemsets Ck, using the apriori-gen() function. This function first joins Lk-1with Lk-1, the joining condition being that the lexicographically ordered first k-2 items are the same. Next,it deletes all those itemsets from the join result that have some (k-1)-subset that is not in Lk-1 yielding Ck.The algorithm now scans the database. For each transaction, it determines which of the candidates in Ckare contained in the transaction using a hash-tree data structure and increments the count of thosecandidates . At the end of the pass, Ck is examined to determine which of the candidates arefrequent, yielding Lk. The algorithm terminates when Lk becomes empty.

V. RECOMMENDATION SYSTEM

Recommendation as a social process plays an important role in many applications for consumers, because it is overly expensive for every consumer to learn about all possible alternatives independently. Depending on the specific application setting, a consumer might be a buyer (e.g., in online shopping), an information seeker (e.g., in information retrieval), or an organization searching for certain expertise. In addition, recommendation as a personalized marketing mechanism has recently attracted significant industry interest (e.g., online shopping and advertising). Recommender systems[1,2,3]have been developed to automate the recommendation process. Examples of research prototypes of recommender systems are: *PHOAKS*, *Syskills* ,*Webert* ,*Fab* , and *GroupLens*. These systems recommend various types of Web resources, online news, movies, among others, to potentially interested parties. Largescale commercial applications of the recommender systems can be found at many e-commerce sites, such as *Amazon*, *CDNow*, *Drugstore*, and *MovieFinder*. These commercial systems recommend products to potential consumers based on previous transactions and feedback. They are becoming part of the standard e-business technology that can enhance e-commerce sales by converting browsers to buyers, increasing cross-selling, and building customer loyalty.

One of the most commonly-used and successful recommendation approaches is the collaborative filtering(CF) approach. When predicting the potential interests of a given consumer, such an approach first identifies a set of similar consumers based on past transaction and product feedback information and then makes a prediction based on the observed behavior of these similar consumers.

VI. RESULT AND ANALYSIS

Methods	Limitations
Classification	Classification is used to build up an idea of the type of customer, item, or object by describing multiple attributes to identify a particular class.This techniques is not efficient for recommendation of the products.
Clustering	In the clustering due to some noisy information it may recommend the unrelated products to the customers in E-commerce.
Sequential pattern and prediction	In sequential pattern and prediction the recommendation is Poor scalability and are not suitable for application that need to extract information from a large number of web sources.
Association rule mining	Association (or relation) is probably the better known and most familiar and straightforward data mining technique. Here, we make a simple correlation between two or more items, often of the same type to identify patterns.

VII. CONCLUSION

Here we have studied, how distributed data mining(in a broad sense, DM applied to ecommerce)is applicable to improve the services provided by e-commerce based enterprises.Specifically, we first discussed some popular tools and techniques used in data mining. One of the main tool is Recommendation system connect users with items to "consume" (purchase, view, listen to, etc.) by associating the content of recommended items or the opinions of other individuals with the consuming user's actions or opinions, based on the collaborative filtering to filter the products.Collaborative filtering (CF) is an effective method of recommender systems (RS) has been widely used in online stores.

REFERENCES

[1] *Luo Zhenghua, "Realization Of Individualized Recommendation System On Books Sale"* IEEE 2012 International Conference on Management ofe-Commerce and e-Government. pp.10-13.

[2] Yechun Jiang, Jianxun Liu, Mingdong Tang and Xiaoqing (Frank) Liu "An Effective Web Service Recommendation Method based on Personalized Collaborative Filtering", 2011 IEEE International Conference on Web Services.

[3] Cane Wing-ki Leung, Stephen Chi-fai Chan and Fu-lai Chung "Applying associative retrival technique tonalleviate the sparsity problem in collaborative filtering", 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology.

[4] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005

[5] K. Yu, A. Schwaighofer, V. Tresp, X. Xu, and H.-P. Kriegel, "Probabilistic memory-based collaborative filtering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 1, pp. 56–69, 2004

[6] Xiaoyuan Su and Taghi M. Khoshgoftaar "A Survey of Collaborative Filtering Techniques", Hindawi Publishing Corporation, Advances in Artificial Intelligence Volume 2009.

[7] Jong-Seok Lee, Sigurdur Olafsson, "Two-way cooperative prediction for collaborative filtering recommendations, Expert Systems with Applications: "An International Journal Volume 36 , Issue 3 April 2009

[8] Z.B. Zheng, H. Ma, M.R. Lyu, and I. King, "WSRec: a collaborative filtering based Web service recommendation system," Proc. 7th International Conference on Web Services (ICWS 2009), 2009, pp. 437-444.

[9] S.S. Weng, B.S. Lin, and W.T. Chen, "Using ContextualInformation and Multidimensional Approach For Recommendation," Expert Systems with Applications, 36(2),2009, pp. 1268-1279

[10] Guangping Zhuo, Jingyu Sun and Xueli Yu "A Framework for Multi-Type Recommendations", Eighth International Conference on FuzzySystems and Knowledge Discovery, 2007.

[11] SARWAR, B., KARYPIS, G., KONSTAN, J., AND REIDL, J. 2001"Item-based collaborative filtering recommendation algorithms" In *Proceedings of the 10th International Conference on World Wide Web(WWW'01)*. ACM, New York, NY, pp.285–295.

[12] J. Han, M. Kamber, "*Data Mining: Concepts and Techniques*," The Morgan Kaufmann Series, 2001.

[13] Sehafer JB □ Konstan J □ Riedl J. "Recommender systems in ecommerce." In:Proc. Of the 1st ACMConf. on Electronic Commerce[J]. NewYork:ACMPress,1999:158- 166.

[14]N. S. PAPASPYROU, C. E. SGOUROPOULOU and E. S. SKORDALAKIS, "A Model of Collaborating Agents for Content-Based Electronic Document Filtering," Journal of Intelligent and Robotic Systems 26: 199–213, 1999.

[15]Schafer, J. B., Konstan, J. & Riedl, J. (2001) "E- Commerce Recommendation Applications, Data Mining & Knowledge Discovery", 5, pp. 115 ~ 153

[16] L. Terveen, W. Hill, B. Amento, D. McDonald, and J. Creter. PHOAKS: "A system for sharing recommendations."Communications of the ACM, 40(3):59{62, 1997.

[17] L. H. Ungar and D. P. Foster. "Clustering methods for collaborative _ltering."In Workshop on Recommendation Systems at the 15th National Conference on Arti_cial Intelligence, 1998.

[18] Agrawal, R., Imielinski, T., Swami, A. N. "Mining association rules between sets of items in large databases". In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216, 1993.

[19]Aggarwal, C.C., Wolf,J.L.,Wu,K.,and Yu,P.S.(1999). "Horting hatches an engg:A new graph theoretic approach to collaborative filtering." In preceding of the ACM KDD'99 Conference. San Diego, CA.pp.201-212.

[20] Basu,c.,Hirsh ,H.,and cohen ,W.(1998)."Recommendation as classification: using social and content based information in recommendation" In recommender system workshop '98 .pp.11-15.

[21] J. Basilico and T. Hofmann. "Unifying collaborative and content-based filtering." the21st International Conference on Machine Learning (ICML), 2004.

[22] Chuangguang Huang and Jian Yin "Effective Association Clusters Filtering to Cold-Start Recommendations", 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery.

[23] Mustansar Ali Ghazanfar and Adam Prugel-Bennett," A Scalable, Accurate Hybrid Recommender System", 2010 Third International Conference on Knowledge Discovery and Data Mining.

[24] Ibrahim A. Almosallam and Yi Shang "A New Adaptive Framework for Collaborative Filtering Prediction", *2008 IEEE Congress on Evolutionary Computation (CEC 2008).*

[25]MOBASHER, B. H., DAI, T. L., NAKAGAWA, M., SUN, Y., ANDWILTSHIRE, J. 2000. "Discovery of aggregate usage profiles for web personalization." In *Proceedings of the Workshop on Web Mining for ECommerce—Challenges and Opportunities*.

[26]. H. Kargupta, I. Hamzaoglu, and B. Stafford. "Scalable, distributed data mining using anagent based architecture". In *Proceedings theThird International Conference on theKnowledgeDiscovery and Data Mining, AAAIPress, Menlo Park, California*, pages 211–214, 1997.

[27]. H. Kargupta and K. Sivakumar. Existential pleasures of distributed data mining. *DataMining:Next Generation Challenges andFuture Directions*, pages 1–25, 2004.

[28]. M. Klusch, S. Lodi, and G. Moro. The role of agents in distributed data mining: Issues andbenefits. In *Proceedings of the*

2003IEEE/WIC International Conference onIntelligent AgentTechnology (IAT 2003),pages 211–217, 2003.

[29]. Mohammed J. Zaki. Parallel and distributed association mining: A survey. *IEEEConcurrency*, 7(4):14–25, 1999.

[30] H.Kargupta, I.Hamzaoglu and B.Stafford "Scalable, Distributed Data Mining An Agent Based Application".Proceedings ofKnowledge Discovery And Data Mining,August, 1997

[31]Byung H. Park and Hillol Kargupta.Distributed data mining: Algorithms,systems, and applications. In Nong Ye,editor, *Data Mining Handbook*, pages 341–358. Lawrence Earlbaum Associates, 2002.

[32] Aronis, J. Kulluri, V., Provost, F., & Buchanan, B. (1997). The WoRLD: Knowledge discovery and multipledistributed databases. in Proceeding of florida artificial intellegence research symposium (FLAIRS-97).

[33]Ashrafi, M.Z., Taniar, D. & Smith, K.(2004). ODAM: An Optimized Distributed Association Rule MiningAlgorithm. IEEE Distributed Systems Online, 5(3).

[34]Cheung, D.W., Ng, V., Fu, A.W. & Fu, Y. (1996). Efficient Mining of Association Rules in DistributedDatabases. IEEE Transactions on Knowledge and Data Engineering, 8(6), 911-922.

[35].Krishnaswamy.s,Zaslavsky,A.,&Loke,S.(2000),"An architecture to support distributed data mining services in E-commerce environment",In second international workshop on advance issues of E-commerce and web based information system(wecuris 2000),Milpitas,CA

[36].Lam,W.,&sergre,A.M(1997)." Distributed data mining(DDM) of probabilistic knowledge". In proceeding of the 17th international conference on distributed computing systems(pp. 178-185) Washington: IEEE computer society press.