JAVA/J2EE PROJECT ABSTRACTS

(Big Data, Cloud Computing, Networking, Network-Security, Mobile Computing, Wireless Sensor Network, Datamining, Webmining, Artificial Intelligence, Vanet, Ad-Hoc Network)

IEEE 2016 / 2015 /2014 / 2013 Papers

DATAMINING

1. Mining Partially-Ordered Sequential Rules Common to Multiple Sequences

Sequential rule mining is an important data mining problem with multiple applications. An important limitation of algorithms for mining sequential rules common to multiple sequences is that rules are very specific and therefore many similar rules may represent the same situation. This can cause three major problems: (1) similar rules can be rated quite differently, (2) rules may not be found because they are individually considered uninteresting, and (3) rules that are too specific are less likely to be used for making predictions. To address these issues, we explore the idea of mining "partially-ordered sequential rules" (POSR), a more general form of sequential rules such that items in the antecedent and the consequent of each rule are unordered. To mine POSR, we propose the RuleGrowth algorithm, which is efficient and easily extendable. In particular, we present an extension (TRuleGrowth) that accepts a sliding-window constraint to find rules occurring within a maximum amount of time. A performance study with four real-life datasets show that RuleGrowth and TRuleGrowth have excellent performance and scalability compared to baseline algorithms and that the number of rules discovered can be several orders of magnitude smaller when the sliding-window constraint is applied. Furthermore, we also report results from a real application showing that POSR can provide a much higher prediction accuracy than regular sequential rules for sequence prediction.

2. Mining High Utility Patterns in One Phase without Generating Candidates

Utility mining is a new development of data mining technology. Among utility mining problems, utility mining with the itemset share framework is a hard one as no anti-monotonicity property holds with the interestingness measure. Prior works on this problem all employ a two-phase, candidate generation approach with one exception that is however inefficient and not scalable with large databases. The two-phase approach suffers from scalability issue due to the huge number of candidates. This paper proposes a novel algorithm that finds high utility patterns in a single phase without generating candidates. The novelties lie in a high utility pattern growth approach, a lookahead strategy, and a linear data structure. Concretely, our pattern growth approach is to search a reverse set enumeration tree and to prune search space by utility upper bounding. We also look ahead to identify high utility patterns without enumeration by a closure property and a singleton property. Our linear data structure enables us to compute a tight bound for powerful pruning and to directly identify high utility patterns in an efficient and scalable way, which targets the root cause with prior algorithms. Extensive experiments on sparse and dense, synthetic and real world data suggest that our algorithm is up to 1 to 3 orders of magnitude more efficient and is more scalable than the state-of-the-art algorithms.

3. TASC: Topic- Adaptive Sentiment Classification on Dynamic Tweets

Sentiment classification is a topic-sensitive task, i.e., a classifier trained from one topic will perform worse on another. This is especially a problem for the tweets sentiment analysis. Since the topics in Twitter are very diverse, it is impossible to train a universal classifier for all topics. Moreover, compared to product review, Twitter lacks data labeling and a rating mechanism to acquire sentiment labels. The extremely sparse text of tweets also brings down the performance of a sentiment classifier. In this paper, we propose a semi-supervised topic-adaptive sentiment classification (TASC) model, which starts with a classifier built on common features and mixed labeled data from various topics. It minimizes the hinge loss to adapt to unlabeled data and features including topic-related sentiment words, authors' sentiments and sentiment connections derived from "@" mentions of tweets, named as topic-adaptive features. Text and non-text features are extracted and naturally split into two views for co-training. The TASC learning algorithm updates topic-adaptive features based on the collaborative selection of unlabeled data, which in

turn helps to select more reliable tweets to boost the performance. We also design the adapting model along a timeline (TASC-t) for dynamic tweets. An experiment on 6 topics from published tweet corpuses demonstrates that TASC outperforms other well-known supervised and ensemble classifiers. It also beats those semi-supervised learning methods without feature adaption. Meanwhile, TASC-t can also achieve impressive accuracy and F-score. Finally, with timeline visualization of "river" graph, people can intuitively grasp the ups and downs of sentiments' evolvement, and the intensity by color gradation.

4. A Compendium for Prediction of Success of a Movie Based Upon Different Factors

The success of a movie is uncertain but it is no secret that it is dependent to a large extent upon the level of promotion and also upon the ratings received by the major movie critics. Time and money are valuable to the general audience and hence, they refer to the leading critics when making a decision about whether to watch a particular movie or not. Over 1000 movies on an average are produced per year. Therefore, in order to make the movie profitable, it becomes a matter of concern that the movie succeeds. Due to the low success rate, models and mechanisms to predict reliably the ranking and or box office collections of a movie can risk the business significantly. The current predictive models that are available are based on various factors for assessment of the movie. These include the typical factors such as cast, producer, director etc. or the social factors in form of response of the society on various online platforms. Various stakeholders such as actors, financiers, directors etc. can use these predictions to make more informed decisions.

5. SEGMENTING CUSTOMERS WITH DATA MINING TECHNIQUES

Retail marketers are constantly looking for ways to improve the effectiveness of their campaigns. One way to do this is to target customers with the particular offers most likely to attract them back to the store and to spend more time and money on their next visit. Demographic market segmentation is an approach to segmenting markets. A company divides the larger market into groups based on several defined criteria. Age, gender, marital status, occupation, education and income are among the commonly considered demographics segmentation criteria. A sample case study has been done in order to explain the theory of segmentation applied on a Turkish supermarket chain. The purpose of this case study is to determine dependency on products and shopping habits. Furthermore forecast sales determine the promotions of products and customer profiles. Association rule mining was used as a method for identifying customers buying patterns and as a result customer profiles were determined. Besides association rules, interesting results were found about customer profiles, such as "What items do female customers buy?" or "What do consumers(married and 35-45 aged) prefer mostly?". For instance, female customers purchase feta cheese with a percentage of 60% whereas male customers purchase tomato with a percentage of 46%. Regarding to customers age, 65 and older customers purchase tea with a percentage of 58%, and customers aged between 18- 25 preferred pasta with a percentage of 57%

6. Canopy Clustering Based K Strange Point Detection.

A Theoretical Comparison of Job scheduling Algorithms in Cloud Computing Environment
Cloud computing is a dynamic, scalable and payper-use distributed computing model empowering designers to convey applications amid job designation and storage distribution. Cloud computing encourages to impart a pool of virtualized computer resource empowering designers to convey applications amid job designation and storage distribution. The cloud computing mainly aims to give proficient access to remote and geographically distributed resources. As cloud technology is evolving day by day and confronts numerous challenges, one of them being uncovered is scheduling. Scheduling is basically a set of constructs constructed to have a controlling hand over the order of work to be performed by a computer system. Algorithms are vital to schedule the jobs for execution. Job scheduling algorithms is one of the most challenging hypothetical problems in the cloud computing domain area. Numerous deep investigations have been carried out in the domain of job scheduling of cloud computing. This paper

intends to present the performance comparison analysis of various pre-existing job scheduling algorithms considering various parameters. This paper discusses about cloud computing and its constructs in section (i). In section (ii) job scheduling concept in cloud computing has been elaborated. In section (iii) existing algorithms for job scheduling are discussed, and are compared in a tabulated form with respect to various parameters and lastly section (iv) concludes the paper giving brief summary of the work.

7. Automated Discovery of Small Business Domain Knowledge Using Web Crawling and Data Mining

It has become an era where everything is on the web with ever more chances of data utilization on the web. Still, there are obstacles to make the use of the web efficiently. With too much information, Internet users have often come across information that are not relevant for their use. On top of that, until recently, most of web content have not contained semantic information, posing difficulties for mechanical analysis. The Semantic Web emerged as a way to tackle those poor qualities of the web. Adopting formal languages such as RDF or OWL, the semantic web has made the Internet become more highly available for computer-based analysis. In this study, what we aimed at is building a small business knowledge base to provideuseful information for small business owners for their marketing strategies or dynamic QA systems for their restaurant recommendation services. The knowledge base was built according to the concept of the Semantic Web. To build the knowledge base, first, it is needed to conduct web crawling from different web sources including social media. However, the crawled data typically come in informal and do not have any semantic information. So we devised text mining techniques to catch useful information from them and generate formal knowledge for the knowledge base.

8. Clustering Data Streams Based on Shared Density Between Micro-Clusters

As more and more applications produce streaming data, clustering data streams has become an important technique for data and knowledge engineering. A typical approach is to summarize the data stream in real-time with an online process into a large number of so called micro-clusters. Micro-clusters represent local density estimates by aggregating the information of many data points in a defined area. On demand, a (modified) conventional clustering algorithm is used in a second offline step to recluster the micro-clusters into larger final clusters. For reclustering, the centers of the micro-clusters are used as pseudo points with the density estimates used as their weights. However, information about density in the area between micro-clusters is not preserved in the online process and reclustering is based on possibly inaccurate assumptions about the distribution of data within and between micro-clusters (e.g., uniform or Gaussian). This paper describes DBSTREAM, the first micro-cluster-based online clustering component that explicitly captures the density between micro-clusters via a shared density graph. The density information in this graph is then exploited for reclustering based on actual density between adjacent micro-clusters. We discuss the space and time complexity of maintaining the shared density graph. Experiments on a wide range of synthetic and real data sets highlight that using shared density improves clustering quality over other popular data stream clustering methods which require the creation of a larger number of smaller micro-clusters to achieve comparable results.

9. Comparitive Analysis of K-Means and Fuzzy C Menas on Thiroid Disease

To recognize in vast restorative and variation groups with unstructured information is dependably a major test furthermore hazard component to exhibit outside world in an organized configuration. To overcome dependably the information ought to be stemmed and sorted in grouped parts. K-means is utilized as a part of the primary methodology of our as DDC. By

taking the mean estimation of age and the groups will be encircled. Be that as it may, the groups are just mean based arrangement so we propose another methodology after K-Means based manufactured FFM. Taken after by that arbitrary markers set to get the achievable consequences of classification furthermore recurrence of appearance regarding allocated irregular scope of interesting qualities to every last existing mix tuple. To accomplish the wanted arrangement of activity we propose another methodology recognized bunching and attainable recurrence with extraordinary result for each procedure. These successions are trailed by non-standard pre-handling like self-cleaning of information and stemming. The pre-handling is finished by semi fluffy system. To accomplish these things of procedure we propose another calculation called DDC (unmistakable relocation for grouping) and UTOF (Unique recurrence result in expandable information. This procedure is absolutely on extensive therapeutic information which is constantly expandable with different new infections. The above calculation takes after a specific new system called doable fluffy mining (FFM) strategy

10. The Classification Techniques on Medical Data to Predict Heart Disease

Analysis of therapeutic information is very test in context of incremental development of properties and different parameters. Once the information is gathering there is specific farthest point to wind up in getting the information as tuples or ascertaining the recurrence of the combinational credits concerning age, ailment, sexual orientation. The most compelling motivation is to push the mysterious maladies in examination. In substantial information related information spotting of the fancied infection information is excessively muddled. So we utilize KNN with Euclidean separation component and choice tree with ordinary and upgraded model concerning given characteristics. In KNN Euclidean separation produced for all tuples and rank will be accommodate new characterization of new set and result created in light of k worth as closest neighbors. Be that as it may, for examination is between both above said as for investigation of time multifaceted nature for order. The principle characterization is done on tremendous and element patch information. So the fundamental arrangement is done on KNN methodology and immediate arrangement trees make up by utilizing NAE approach (typical and improved choice tree structures).

11. Enriched content mining for web applications

In recent years, it has been witnessed that the ever-interesting and upcoming publishing medium is the World Wide Web. Much of the web content is unstructured so gathering and making sense of such data is very tedious. Web servers worldwide generate a vast amount of information on web users' browsing activities. Several researchers have studied these so-called web access log data to better understand and characterize web users. Data can be enriched with information about the content of visited pages and the origin (e.g., geographic, organizational) of the requests. The goal of this project is to analyze user behavior by mining enriched web access log data. The several web usage mining methods for extracting useful features is discussed and employ all these techniques to cluster the users of the domain to study their behaviors comprehensively. The contributions of this thesis are a data enrichment that is content and origin based and a treelike visualization of frequent navigational sequences. This visualization allows for an easily interpretable tree-like view of patterns with highlighted relevant information. The results of this project can be applied on diverse purposes, including marketing, web content advising, (re-)structuring of web sites and severa l other E-business processes, like recommendation and advertiser systems. It also rank the best relevant

documents based on Top K query for effective and efficient data retrieval system. It filters the web documents by providing the relevant content in the search engine result page (SERP)

12. Text Mining-Supported Information Extraction.

Information extraction (IE) and knowledge discovery in databases (KDD) are both useful approaches for discovering information in textual corpora, but they have some deficiencies. Information extraction can identify relevant subsequences of text, but is usually unaware of emerging, previously unknown knowledge and regularities in a text and thus cannot form new facts or new hypotheses. Complementary to information extraction, emerging data mining methods and techniques promise to overcome the deficiencies of information extraction. This research work combines the benefits of both approaches by integrating datamining and information extraction methods. The aim is to provide a new high-quality information extraction methodology and, at the same time, to improve the performance of the underlying extraction system. Consequently, the new methodology should shorten the life cycle of information extraction engineering because information predicted in early extraction phases can be used in further extraction steps, and the extraction rules developed require fewer arduous test-and-debug iterations. Effectiveness and applicability are validated by processing online documents from the areas of eHealth and eRecruitment.

13. Dynamic Query Forms for Database Queries

Modern scientific databases and web databases maintain large and heterogeneous data. These real-world databases contain hundreds or even thousands of relations and attributes. Traditional predefined query forms are not able to satisfy various ad-hoc queries from users on those databases. This paper proposes DQF, a novel database query form interface, which is able to dynamically generate query forms. The essence of DQF is to capture a user's preference and rank query form components, assisting him/her in making decisions. The generation of a query form is an iterative process and is guided by the user. At each iteration, the system automatically generates ranking lists of form components and the user then adds the desired form components into the query form. The ranking of form components is based on the captured user preference. A user can also fill the query form and submit queries to view the query result at each iteration. In this way, a query form could be dynamically refined until the user is satisfied with the query results. We utilize the expected F-measure for measuring the goodness of a query form. A probabilistic model is developed for estimating the goodness of a query form in DQF. Our experimental evaluation and user study demonstrate the effectiveness and efficiency of the system.

14. Secure Data Mining in Cloud using Homomorphic Encryption

With the advancement in technology, industry, ecommerce and research a large amount of complex and pervasive digital data is being generated which is increasing at an exponential rate and often termed as big data. Traditional Data Storage systems are not able to handle Big Data and also analyzing the Big Data becomes a challenge and thus it cannot be handled by traditional analytic tools. Cloud Computing can resolve the problem of handling, storage and analyzing the Big Data as it distributes the big data within the cloudlets. No doubt, Cloud Computing is the best answer available to the problem of Big Data storage and its analyses but having said that, there is always a potential risk to the security of Big Data storage in Cloud Computing, which needs to be addressed. Data Privacy is one of the major issues while storing the Big Data in a Cloud environment. Data Mining based attacks, a major threat to the data, allows an adversary or an unauthorized user to infer valuable and sensitive information by analyzing the results

generated from computation performed on the raw data. This thesis proposes a secure k-means data mining approach assuming the data to be distributed among different hosts preserving the privacy of the data. The approach is able to maintain the correctness and validity of the existing k-means to generate the final results even in the distributed environment.

15. RuleGrowth: Mining Sequential Rules Common to Several Sequences by Pattern-Growth

Mining sequential rules from large databases is an important topic in data mining fields with wide applications. Most of the relevant studies focused on finding sequential rules appearing in a single sequence of events and the mining task dealing with multiple sequences were far less explored. In this paper, we present RuleGrowth, a novel algorithm for mining sequential rules common to several sequences. Unlike other algorithms, RuleGrowth uses a pattern-growth approach for discovering sequential rules such that it can be much more efficient and scalable. We present a comparison of RuleGrowth's performance with current algorithms for three public datasets. The experimental results show that RuleGrowth clearly outperforms current algorithms for all three datasets under low support and confidence threshold and has a much better scalability.

16. Secure Mining of Association Rules in Horizontally Distributed Databases

We propose a protocol for secure mining of association rules in horizontally distributed databases. The current leading protocol is that of Kantarcioglu and Clifton [18]. Our protocol, like theirs, is based on the Fast Distributed Mining (FDM) algorithm of Cheung et al. [8], which is an unsecured distributed version of the Apriori algorithm. The main ingredients in our protocol are two novel secure multi-party algorithms—one that computes the union of private subsets that each of the interacting players hold, and another that tests the inclusion of an element held by one player in a subset held by another. Our protocol offers enhanced privacy with respect to the protocol in [18]. In addition, it is simpler and is significantly more efficient in terms of communication rounds, communication cost and computational cost

17. A Privacy Leakage Upper-Bound Constraint Based Approach for Cost-Effective Privacy Preserving of Intermediate Datasets In cloud

Cloud computing provides massive computation power and storage capacity which enable users to deploy computation and data intensive applications without infrastructure investment. Along the processing of such applications, a large volume of intermediate datasets will be generated, and often stored to save the cost of re-computing them. However, preserving the privacy of intermediate datasets becomes a challenging problem because adversaries may recover privacy-sensitive information by analyzing multiple intermediate datasets. Encrypting ALL datasets in cloud is widely adopted in existing approaches to address this challenge. But we argue that encrypting all intermediate datasets are neither efficient nor cost-effective because it is very time consuming and costly for data-intensive applications to en/decrypt datasets frequently while performing any operation on them. In this paper, we propose a novel upper-bound privacy leakage constraint based approach to identify which intermediate datasets need to be encrypted and which do not, so that privacy-preserving cost can be saved while the privacy requirements of data holders can still be satisfied. Evaluation results demonstrate that the privacy-preserving cost of intermediate datasets can be significantly reduced with our approach over existing ones where all datasets are encrypted.

18. Automatic Medical Disease Treatment System Using Datamining

In our proposed system is identifying reliable information in the medical domain stand as building blocks for a healthcare system that is up-todate with the latest discoveries. By using the tools such as NLP, ML techniques. In this research, focus on diseases and treatment information, and the relation that exists between these two entities. The main goal of this research is to identify the disease name with the symptoms specified and extract the sentence from the article and get the Relation that exists between Disease- Treatment and classify the information into cure,

prevent, side effect to the user.This electronic document is a "live" template. The various components of your paper [title, text, heads, etc.] are already defined on the style sheet, as illustrated by the portions given in this document.

19. Efficient Algorithms for Mining High Utility Itemsets from Transactional Databases

Mining high utility itemsets from a transactional database refers to the discovery of itemsets with high utility like profits. Although a number of relevant algorithms have been proposed in recent years, they incur the problem of producing a large number of candidate itemsets for high utility itemsets. Such a large number of candidate itemsets degrades the mining performance in terms of execution time and space requirement. The situation may become worse when the database contains lots of long transactions or long high utility itemsets. In this paper, we propose two algorithms, namely utility pattern growth (UP-Growth) and UP-Growth+, for mining high utility itemsets with a set of effective strategies for pruning candidate itemsets. The information of high utility itemsets is maintained in a tree-based data structure named utility pattern tree (UP-Tree) such that candidate itemsets can be generated efficiently with only two scans of database. The performance of UP-Growth and UP-Growth+ is compared with the state-of-the-art algorithms on many types of both real and synthetic data sets. Experimental results show that the proposed algorithms, especially UPGrowth+, not only reduce the number of candidates effectively but also outperform other algorithms substantially in terms of runtime,especially when databases contain lots of long transactions.

20. Automatic Itinerary Planning for Traveling Services

Creating an efficient and economic trip plan is the most annoying job for a backpack traveler. Although travel agency can provide some predefined itineraries, they are not tailored for each specific customer. Previous efforts address the problem by providing an automatic itinerary planning service, which organizes the points-of-interests (POIs) into a customized itinerary. Because the search space of all possible itineraries is too costly to fully explore, to simplify the complexity, most work assume that user's trip is limited to some important POIs and will complete within one day. To address the above limitation, in this paper, we design a more general itinerary planning service, which generates multiday itineraries for the users. In our service, all POIs are considered and ranked based on the users' preference. The problem of searching the optimal itinerary is a team orienteering problem (TOP), a well-known NP-complete problem. To reduce the processing cost, a two-stage planning scheme is proposed. In its preprocessing stage, single-day itineraries are precomputed via the MapReduce jobs. In its online stage, an approximate search algorithm is used to combine the single day itineraries. In this way, we transfer the TOP problem with no polynomial approximation into another NP-complete problem (set-packing problem) with good approximate algorithms. Experiments on real data sets show that our approach can generate high-quality itineraries efficiently.

21. Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance

The vast majority of existing approaches to opinion feature extraction rely on mining patterns only from a single review corpus, ignoring the nontrivial disparities in word distributional characteristics of opinion features across different corpora. In this paper, we propose a novel method to identify opinion features from online reviews by exploiting the difference in opinion feature statistics across two corpora, one domain-specific corpus (i.e., the given review corpus) and one domain-independent corpus (i.e., the contrasting corpus). We capture this disparity via a measure called domain relevance (DR), which characterizes the relevance of a term to a text collection. We first extract a list of candidate opinion features from the domain review corpus by defining a set of syntactic dependence rules. For each extracted candidate feature, we then estimate its intrinsic-domain relevance (IDR) and extrinsic-domain relevance (EDR) scores on the domain-dependent and domain-independent corpora, respectively. Candidate features that are less generic (EDR score less than a threshold) and more domain-specific

(IDR score greater than another threshold) are then confirmed as opinion features. We call this interval thresholding approach the intrinsic and extrinsic domain relevance (IEDR) criterion. Experimental results on two real-world review domains show the proposed IEDR approach to outperform several other well-established methods in identifying opinion features.

22. Keyword Query Routing

Keyword search is an intuitive paradigm for searching linked data sources on the web. We propose to route keywords only to relevant sources to reduce the high cost of processing keyword search queries over all sources. We propose a novel method for computing top-k routing plans based on their potentials to contain results for a given keyword query. We employ a keyword-element relationship summary that compactly represents relationships between keywords and the data elements mentioning them. A multilevel scoring mechanism is proposed for computing the relevance of routing plans based on scores at the level of keywords, data elements, element sets, and subgraphs that connect these elements. Experiments carried out using 150 publicly available sources on the web showed that valid plans (precision@1 of 0.92) that are highly relevant (mean reciprocal rank of 0.89) can be computed in 1 second on average on a single PC. Further, we show routing greatly helps to improve the performance of keyword search, without compromising its result quality.

23. The Role of Apriori Algorithm for Finding the Association Rules in Data Mining

Now a day's Data mining has a lot of e-Commerce applications. The key problem is how to find useful hidden patterns for better business applications in the retail sector. For the solution of these problems, The Apriori algorithm is one of the most popular data mining approach for finding frequent item sets from a transaction dataset and derive association rules. Rules are the discovered knowledge from the data base. Finding frequent item set (item sets with frequency larger than or equal to a user specified minimum support) is not trivial because of its combinatorial explosion. Once frequent item sets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence. The paper illustrating apriori algorithm on simulated database and finds the association rules on different confidence value

24. Efficient Mining of Both Positive and Negative Association Rules

This paper presents an efficient method for mining both positive and negative association rules in databases. The method extends traditional associations to include association rules of forms A) :B, :A) B, and :A) :B, which indicate negative associations between itemsets. With a pruning strategy and an interestingness measure, our method scales to large databases. The method has been evaluated using both synthetic and real-world databases, and our experimental results demonstrate its effectiveness and efficiency. Categories and Subject Descriptors: I.2.6

25. RuleGrowth: Mining Sequential Rules Common to Several Sequences by Pattern-Growth

Mining sequential rules from large databases is an important topic indata miningfields with wide applications. Most of the relevant studies focused on finding sequential rules appearing in a single sequence of events and the mining task dealing with multiple sequences were far less explored. In this paper, we present RuleGrowth, a novel algorithm for mining sequential rules common to several sequences. Unlike other algorithms, RuleGrowth uses a pattern-growth approach for discovering sequential rules such that it can be much more efficient and scalable. We present a comparison of RuleGrowth's performance with current algorithms for three public datasets. The experimental results show that RuleGrowth clearly outperforms current algorithms for all three datasets under low support and confidence threshold and has a much better scalability

26. Secure Mining of Association Rules in Horizontally Distributed Databases

JAVA/J2EE PROJECT ABSTRACTS

(Big Data, Cloud Computing, Networking, Network-Security, Mobile Computing, Wireless Sensor Network, Datamining, Webmining, Artificial Intelligence, Vanet, Ad-Hoc Network**)**

We propose a protocol for secure mining of association rules in horizontally distributed databases. The current leading protocol is that of Kantarcioglu and Clifton [18]. Our protocol, like theirs, is based on the Fast Distributed Mining (FDM) algorithm of Cheung et al. [8], which is an unsecured distributed version of the Apriori algorithm. The main ingredients in our protocol are two novel secure multi-party algorithms—one that computes the union of private subsets that each of the interacting players hold, and another that tests the inclusion of an element held by one player in a subset held by another. Our protocol offers enhanced privacy with respect to the protocol in [18]. In addition, it is simpler and is significantly more efficient in terms of communication rounds, communication cost and computational cost