Personalized market basket prediction with temporal annotated recurring sequences

A.KANIMOZHI

M.Phil scholar

Rathnavel Subramaniam college of arts and science.

ABSTRACT

The emergence of the business-to-customer (B2C) markets has resulted in various studies on developing and improving customer retention and profit enhancement. This is mainly due to the retail business becoming increasingly competitive with costs being driven down by new and existing competitors. In general, consumer markets have several characteristics such as repeat buying over the relevant time interval, a large number of customers, and a wealth of information detailing past customer purchases

The provision of customized service to the customers is vital for a company to establish long lasting and pleasant relationship with consumers. It has also been observed that keeping old customers generates more profit than attracting new ones. So, customer retention is a big factor too. So, there is always a trade-off between customer benefits and transaction costs, which has to be optimized by the managers.

The purpose of this thesis is to study, implement and analyze various Data-mining tools and techniques and then do an analysis of the sample / raw data to obtain a meaningful interpretation. Some of the data mining algorithms I have used, are a vector quantization based clustering algorithm, and then an 'Apriori' based Association rule mining algorithm. The first one is aimed at a meaningful segregation of the various customers based on their RFM values, while the latter algorithm tries to find out relationships and patterns among the purchases made by the customer, over several transactions.

CHAPTER – I

INTRODUCTION

1.1. INTRODUCTION TO DATA MINING

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?"

1.1.1. The Scope of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database — for example, finding linked products in gigabytes of store scanner data — and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides.

1.2. CONSUMER BEHAVIOUR USING DATA MINING

Various studies on consumer purchasing behaviors have been presented and used in real problems. Data mining techniques are expected to be a more effective tool for analyzing consumer behaviors. However, the data mining method has disadvantages as well as advantages. Therefore, it is important to select appropriate techniques to mine databases. The objective of this paper is to know consumer behavior, his psychological condition at the time of purchase and how suitable data mining method apply to improve conventional method. Moreover, in an experiment, association rule is employed to mine rules for trusted customers using sales data in a super market industry

Consumer behavior means the study of individuals, groups or organizations about their process of selecting, securing, using and disposing the products, services, experiences or ideas to satisfy needs and the impact of these process on the consumer and the society. Behavior concerns either with the individual or the group (e.g. In college friends influence what kind of clothes a person should wants to wears) or a firm (peoples working in firm make decision as to which products the firm should use.) The use of product is often so important to the marketer because this may influence how a product is best positioned or how we can encourage increased consumption. Consumer behavior involves services and ideas as well as tangible products.

In “Market basket analysis in multiple store environments” the author Yen-ling chen, Kwei Tang, Ren-Jie shen, Ya-han Hu find out that there are two main problem in using the existing methods which are used in a multi-store environment. The first is caused by the temporal nature of purchasing pattern. An apparent example is seasonal products. The second problem is associated with finding common association pattern in subset of store. To overcome this problem the authors develop an Apriori like algorithm for automatically extracting association rules in multi-store environment. Various studies on consumer purchasing behavior have been presented and used in real Problem. Data mining techniques are expected to be more effective tool for analyzing consumer behavior. However the data mining methods has disadvantages as well as advantages.

Over the years Data mining (DM) can used to understand the consumer buying behavior using various techniques. Data mining has gradually increases many folds and today it is a giant 100 billion dollar industry. In data mining world every activity of a consumer in a supermarket is treated as a byte of data. How the consumer spends, which day what time normally he/she does the shopping, what they buy most often, how much they buy, in that locality etc. All this data which is gathered somewhere at the backend about which a consumer is not even aware and there is a big industry which is slicing & dicing this data & selling it at a premium price.

Data mining is the method of analyzing data from different angle or perspective and collecting it to get useful information that can be used to increase revenue costs or both, DM allows backend processors to analyze data from many different dimensions, categories it & summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozen of fields in large relational databases .Data mining is primarily used day by day comprise with a strong consumer focus retail, financial , communication & marketing organizations. It enables this companies to determine relationships among internal factors such as price, product positioning or staff skills & external factors such as economic indicators, competition & customer demographics. Most of the time the data is used to analyze the pattern or shopping habits of consumers like in festive season which product sells more. What are the association between these product? Is there any pattern in this habit? If data show some common theme then stores management arrange that product accordingly. e.g. If management arrange electronic product like television, LCDs, Tape recorders, Mobiles etc. with attractive schemes in the front row in festive season. And also arrange the similar items which customer tends to buy along with these product. To make more profit stores will not run any discount or special offers on the products on busy days. Yet another area commonly traced is the weekly shopping habit of the customer, What products they buy and of what quality. This information can be used for stocking purposes and handle the inventory cost. Likewise there are many other aspects in which this data analysis is leading to better consumer satisfaction. For monthly analysis about the certain product demand i.e. buying in the start of the month and buying at the end of the month? The people have money to spent in the starting of the month and at the end of the month people spend less. In the vacation of the school and at the starting of the school the requirement of certain commodity is increase. So to maintain the inventory and also to increase sell in this period. It is necessary to grab this opportunity of consumer needs. In all these cases the data which is collected from different sources. Some of these are operational or transactional data such as sales, cost, inventory, payroll and accounting. Non operational data such as industry sales, forecast data and macro economic data. Meta-data Data about data itself, such as logical database design or data dictionary definition.

1.3. MARKET BASKET ANALYSIS

A market basket analysis or recommendation engine is what is behind all these recommendations we get when we go shopping online or whenever we receive targeted advertising. The underlying engine collects information about people’s habits and knows that if people buy pasta and wine, they are usually also interested in pasta sauces. So, the next time you go to the supermarket and buy pasta and wine, be ready to get a recommendation for some pasta sauce!

Market Basket Analysis is one of the key systems utilized by substantial retailers to reveal relationship between things. With the assistance of market Basket Analysis, the computer will be capable to discover purchasing patterns independent from anyone else without being advised what patterns to search for. Seeing such patterns include various applications inside disciplines like classification administration and advertising and can be utilized to:

- Optimize store layout (space management) by putting items that are often bought together in near proximity to each other – and by allocating shelf space not only according to sales performance of the product itself, but also according to sales of other items the product drives
- Increase cross-selling and up-selling by understanding and leveraging the roles of product categories (e.g. basket drivers, traffic drivers etc.)
- Deliver targeted online marketing campaigns to customers based on their purchasing behavior
- Drive recommendation engines for online retailers
- Deliver targeted online marketing campaigns to customers based on their purchasing behavior

Customer Relationship Management (CRM) is an ocean of data from sales and marketing, customer support, and product development. By using big data & Machine Learning CRM can add value to the organization. It can improve ROI and drive better outcomes. Machine learning can optimize how you oversee, comprehend, and serve clients—both at the individual client level and across your entire client base. It is made of 3 phases:

- Analyzing the past to understand what actions and data led to great outcomes, such as high customer satisfaction;
- Interpreting each new customer interaction and making recommendations on the next best action to influence a successful outcome; and

Continually updating its "learning" based on the most recent set of outcomes to remain relevant without the need for manual changes and inputs.

1.4. ORGANIZATION OF THE THESIS

The remainder of this thesis is organized into 5 chapters.

Chapter 1 describes the introduction to the study, such as what is data mining and the need of data mining usages.

Chapter 2 describes the background study on methods related to market basket analysis using association rule mining.

Chapter 3 provides the proposed methods with their operations and contributions of this thesis.

Chapter 4 provides the experimental evaluation on public benchmarks, and the corresponding critical discussion.

Chapter 5 deals with the conclusion of the thesis with the future work.

1.5. OBJECTIVE OF THE RESEARCH

The thesis aims to market basket prediction at over-coming the main limitations of existing methods. To this purpose, we propose a Association rule mining which combines ideas underlying sequential and pattern-based recommenders. The approach consists of two main components. Association rule mining from the customer’s purchase history, i.e., sequential recurring patterns able to capture the customer’s purchasing habits.

CHAPTER – II

LITERATURE SURVEY

2.1. USING ASSOCIATION RULES FOR PRODUCT ASSORTMENT DECISIONS: A CASE STUDY

INTRODUCTION

In the past, retailers saw their job as one of buying products and putting them out for sale to the public. If the products were sold, more were ordered. If they did not sell, they were disposed of. Blischok [7] describes retailing in this model as a product oriented business, where talented merchants could tell by the look and feel of an item whether or not it was a winner. In order to be successful, retailing today can no longer be just a product-oriented business. According to Blischok, it must be a customer-oriented business and superior customer service comes from superior knowledge of the customer. It is defined as the understanding of all customer's purchasing behaviour as revealed through his or her sales transactions, i.e. market basket analysis. Currently, the gradual availability of cheaper and better information technology has in many retail organisations resulted in an abundance of sales data. Hedberg [17] mentions the American supermarket chain ‘Wal-Mart’ which stores about 20 million sales transactions per day. This explosive growth of data leads to a situation in which retailers today find it increasingly difficult to obtain the right information, since traditional methods of data analysis cannot deal effectively with such huge volumes of data. This is where knowledge discovery in databases (KDD) comes into play. Today, among the most popular techniques in KDD, is the extraction of association rules from large databases. While many researchers have significantly contributed to the development of efficient association rule algorithms [1-3, 10, 21, 26], literature on the use of this technique in concrete real-world applications remains rather limited [4, 5, 25]. Nevertheless, the widespread acceptance of association rules as a valuable technique to solve real business problems will largely depend on the successful application of this technique on real-world data. Moreover, it has been claimed recently [18] that the utility of extracted patterns (such as association rules) in decision-making can only be addressed within the microeconomic framework of the enterprise. This means that a pattern in the data is interesting only to the extent in which it can be used in the decision-making process of the enterprise to increase utility. In this paper, we tackle the problem of product assortment analysis and introduce a concrete microeconomic integer programming model for product selection (PROFSET1) based on the use of frequent item sets. We demonstrate its effectiveness on real-world sales transaction data obtained from a fully-automated convenience store.

2.2. ASSOCIATION RULES AND DATA MINING IN HOSPITAL INFECTION CONTROL AND PUBLIC HEALTH SURVEILLANCE

The authors consider the problem of identifying new, unexpected, and interesting patterns in hospital infection control and public health surveillance data and present a new data analysis process and system based on association rules to address this problem.

The authors first illustrate the need for automated pattern discovery and data mining in hospital infection control and public health surveillance. Next, they define association rules, explain how those rules can be used in surveillance, and present a novel process and system--the Data Mining Surveillance System (DMSS)--that utilize association rules to identify new and interesting patterns in surveillance data.

Experimental results were obtained using DMSS to analyze Pseudomonas aeruginosa infection control data collected over one year (1996) at University of Alabama at Birmingham Hospital. Experiments using one-, three-, and six-month time partitions yielded 34, 57, and 28 statistically significant events, respectively. Although not all statistically significant events are clinically significant, a subset of events generated in each analysis indicated potentially significant shifts in the occurrence of infection or antimicrobial resistance patterns of P. aeruginosa.

The new process and system are efficient and effective in identifying new, unexpected, and interesting patterns in surveillance data. The clinical relevance and utility of this process await the results of prospective studies currently in progress.

2.3. EFFECTIVE PERSONALIZATION BASED ON ASSOCIATION RULE DISCOVERY FROM WEB USAGE DATA

One of the most successful and widely used technologies for building personalization and recommendation systems is collaborative filtering (CF) [20]. Given a target user's record of activity, CF-based techniques, such as the k-NearestNeighbor (kNN) approach, compare that record with the historical records of other users in order to find the top k users who have similar tastes or interests. The mapping of a visitor record to its neighborhood could be based on similar ity in ratings of items, access to similar pages, or purchase of similar items. The identified neighborhood is then used to recommend items not already accessed or purchased by the active user. The CF-based techniques suffer from some well-known limitations [17]. For the most part these limitations are related to the scalability and efficiency of the kNN approach. Essentially, kNN requires that the neighborhood formation phase be performed as an online process, and for very large data sets this may lead to unacceptable latency for providing recommendations. A number of optimization strategies have been proposed and employed to remedy this shortcoming [3, 18]. These strategies include similarity indexing and dimensionality reduction to reduce real-time search costs. The challenge in designing effective Web personalization systems is to improve the scalability of collaborative filtering through offline pattern discovery, while maintaining or improving the overall recommendation effectiveness. Furthermore, the effectiveness of the system must be measured in terms of both coverage and accuracy (precision) of the produced recommendations. Precision measures the degree to which the recommendation engine produces accurate recommendations. On the other hand, coverage measures the ability of the recommendation engine to produce all of the pageviews that are likely to be visited by the user. Both of these measures are essential in evaluating the effectiveness of recommender systems. For example, in e-commerce domain, low precision can easily lead to angry or frustrated users (who receive inaccurate recommendations) while low coverage will result in the site missing cross-sell or up-sell recommendations at critical junctures in users navigation through the site. In recent years there has been an increasing interest and a growing body of work in Web usage mining [16] as an underlying approach to capturing and modeling Web user behavioral patterns and for deriving e-business intelligence. Web usage mining techniques such as clustering that rely on offline pattern discovery from user transactions can be used to improve the scalability of collaborative filtering. For example, previous work such as [9, 10, 12] have considered automatic personalization based on clustering of user transactions and pageviews. However, this is often at the cost of reduced recommendation accuracy. One solution to improve accuracy is presented by [11] using preprocessing techniques such as normalization and significance filtering. Another way is to consider ordering information in personalization. Comparing with non-sequential patterns such as clusters and association rules, sequential patterns contain more precise information about user’s navigational behavior. The use of navigational sequential patterns for predictive user modeling has been extensively studied [5, 14, 19]. The primary focus of all of these studies has been on prefetching of Web pages (i.e., predicting a user’s next access to a page) to improve server performance or network latency. In the context of personalization, however, the narrow focus on navigational sequences often leads to very low recommendation coverage making such techniques less effective as the basis for recommender systems. Some recent studies have considered the use of association rule mining [2, 15] in recommender systems [6, 7, 17]. For the most part, however, these studies have relied on discovering all association rules prior to generating recommendations (thus requiring search among all rules during the recommendation phase) or on real-time generation of association rules from a subset of transactions within a current user’s neighborhood. There has also been little focus on the impact of factors such as the support threshold or the size of user history on the effectiveness of recommendations. In this paper we present a scalable framework for recommender systems using association rule mining from clickstream data. Specifically, we present a data structure for storing the discovered frequent itemsets which is especially suitable for recommender systems. Our recommendation algorithm utilizes this data structure to produce recommendations efficiently in real-time, without the need to generate all association rules from frequent itemsets. Furthermore, through detailed experimental evaluation we show that by using multiple support levels for diferent types of pageviews and varying sized user histories, our framework can overcome some of the shortcomings of recommender systems based on association rules (e.g., low coverage resulting from high support thresholds or larger user histories, and reduced accuracy due to the sparse nature of the data). In fact, we show that the proposed framework can achieve better overall recommendation effectiveness than direct approaches such as the kNN technique in terms of coverage and accuracy.

2.4 MINING GENE EXPRESSION DATABASES FOR ASSOCIATION RULES

Gene expression data, both at the transcript level and at the protein level, can be a valuable tool in the understanding of genes, biological networks, and cellular states. One goal in analyzing expression data is to try to determine how the expression of any particular gene might affect the expression of other genes; the genes involved in this case could belong to the same gene network. By a gene network, we mean a set of genes being expressed together in a non-random pattern. Another goal of expression data analysis is to try to determine what genes are expressed as a result of certain cellular conditions, e.g. what genes are expressed in diseased cells that are not expressed in healthy cells. While early experiments using microarrays profiled only a few samples, more recent experiments profile on the order of dozens or even hundreds of samples, allowing for a more robust statistical analysis of the data. In the near future, data sets containing thousands of samples should become available. As gene expression data sets become larger and larger, spreadsheets will become less and less of an adequate tool for doing analysis (as a single worksheet in Excel can hold no more than 256 columns), and data mining techniques using large databases should find more and more use in analyzing expression data. Many clustering techniques for grouping genes based on similar expression profiles have been explored (Eisen et al., 1998; Tavazoie et al., 1999; Tamayo et al., 1999). One common data mining technique, different from clustering, for finding and describing relationships between different items in a large data set is to look for association rules in the data. An association rule has the form LHS$\Rightarrow$RHS, where LHS and RHS are sets of items, the RHS set being likely to occur whenever the LHS set occurs. Association rules are used widely in the retail industry under the name ‘market basket analysis’. Association rules have been used as well to mine medical record data (Doddi et al., 2001; Stilou et al., 2001). In market basket analysis, an association rule represents a set of items that are likely to be purchased together; for example, the rule {cereal}$\Rightarrow${milk, juice} would state that whenever a customer purchases cereal, he or she is likely to purchase both milk and juice as well in the same transaction. In the analysis of gene expression data, the items in an association rule can represent genes that are strongly expressed or repressed, as well as relevant facts describing the cellular environment of the genes (e.g. a diagnosis for a tumor sample that was profiled, or a drug treatment given to cells in the sample before profiling). An example of an association rule mined from expression data might be {cancer}$\Rightarrow${gene A↑, gene B↓, gene C↑}, meaning that, for the data set that was mined, in most profile experiments where the cells used were cancerous, gene A was measured as being up (i.e. highly expressed), gene B was down (i.e. highly repressed), and gene C was up, altogether. Public gene expression data sets large enough to mine for association rules and obtain meaningful results are already available. Algorithms for finding rules efficiently have been extensively developed in market basket analysis, and we apply a version of one of these algorithms to mine the compendium of Hughes et al. (2000) of pro- files from 300 diverse mutations and chemical treatments in yeast. We find numerous rules in the data, a cursory analysis of some of which reveals numerous associations between certain genes, many of which make sense biologically, others suggesting new hypotheses that may warrant further investigation. In a data set derived from the yeast data set, but with the expression values for each transcript randomly shifted with respect to the experiments, no rules were found, indicating that very few of the rules mined from the actual data set are likely to have existed in the data by chance.

2.5 CPAR: Classification based on predictive association Rules

In recent years, a new approach called associative classification [7, 6] is proposed to integrate association rule mining [1] and classification. It uses association rule mining algorithm, such as Apriori [1] or FPgrowth [5], to generate the complete set of association rules. Then it selects a small set of high quality rules and uses this rule set for prediction. The experiments in [7, 6] show that this approach achieves higher accuracy than traditional classification approaches such as C4.5 [8]. However, associative classification suffers from efficiency due to the facts that it often generates a very large number of rules in association rule mining, and also it takes efforts to select high quality rules from among them. In this paper, we propose a novel approach called CPAR (Classification based on Predictive Association Rules). CPAR inherits the basic idea of FOIL [9] in rule generation and integrates the features of associative classification in predictive rule analysis. In comparison with associative classification, CPAR has the following advantages: (1) CPAR generates a much smaller set of high-quality predictive rules directly from the dataset; (2) to avoid generating redundant rules, CPAR generates each rule by considering the set of “alreadygenerated” rules; and (3) when predicting the class label of an example, CPAR uses the best k rules that this example satisfies. Moreover, CPAR employs the following features to further improve its accuracy and efficiency: (1) CPAR uses dynamic programming to avoid repeated calculation in rule generation; and (2) when generating rules, instead of selecting only the best literal, all the close-to-the-best literals are selected so that important rules will not be missed. CPAR generates a smaller set of rules, with higher quality and lower redundancy in comparison with associative classification. As a result, CPAR is much more time-efficient in both rule generation and prediction but achieves as high accuracy as associative classification.

2.6 FAST ALGORITHMS FOR MINING ASSOCIATION RULES

We consider the problem of discovering association rules between items in a large database of sales transactions. We present two new algorithms for solving this problem that are fundamentally different from the known algorithms. Empirical evaluation shows that these algorithms outperform the known algorithms by factors ranging from three for small problems to more than an order of magnitude for large problems. We also show how the best features of the two proposed algorithms can be combined into a hybrid algorithm, called AprioriHybrid. Scale-up experiments show that Apriori Hybrid scales linearly with the number of transactions.

Apriori Hybrid also has excellent scale-up properties with respect to the transaction size and the number of items in the database.

2.7 SCALABLE ALGORITHMS FOR ASSOCIATION MINING

Association rule discovery has emerged as an important problem in knowledge discovery and data mining. The association mining task consists of identifying the frequent itemsets and then, forming conditional implication rules among them. In this paper, we present efficient algorithms for the discovery of frequent itemsets which forms the compute intensive phase of the task. The algorithms utilize the structural properties of frequent itemsets to facilitate fast discovery. The items are organized into a subset lattice search space, which is decomposed into small independent chunks or sublattices, which can be solved in memory. Efficient lattice traversal techniques are presented which quickly identify all the long frequent itemsets and their subsets if required. We also present the effect of using different database layout schemes combined with the proposed decomposition and traversal techniques. We experimentally compare the new algorithms against the previous approaches, obtaining improvements of more than an order of magnitude for our test databases.

2.8 Mining frequent patterns without candidate generation

In this paper, we develop and integrate the following three techniques in order to solve this problem. First, a novel, compact data structure, called frequent-pattern tree, or FP-tree in short, is constructed, which is extended prefix-tree structure storing crucial, quantitative information about frequent patterns. To ensure that the tree structure is compact and informative, only frequent length-1 items will have nodes in the tree, and the tree nodes are arranged in such a way that more frequently occurring nodes will have better chances of node sharing than less frequently occurring ones. Our experiments show that such a tree is compact, and it is sometimes orders of magnitude smaller than the original database. Subsequent frequent-pattern mining will only need to work on the FP-tree instead of the whole data set. Second, an FP-tree-based pattern-fragment growth mining method is developed, which starts from a frequent length-1 pattern (as an initial suffix pattern), examines only its conditional-pattern base (a “sub-database” which consists of the set of frequent items cooccurring with the suffix pattern), constructs its (conditional) FP-tree, and performs mining recursively with such a tree. The pattern growth is achieved via concatenation of the suffix pattern with the new ones generated from a conditional FP-tree. Since the frequent itemset in any transaction is always encoded in the corresponding path of the frequent-pattern trees, pattern growth ensures the completeness of the result. In this context, our method is not Apriori-like restricted generation-and-test but restricted test only. The major operations of mining are count accumulation and prefix path count adjustment, which are usually much less costly than candidate generation and pattern matching operations performed in most Apriori-like algorithms. Third, the search technique employed in mining is a partitioning-based, divide-and conquer method rather than Apriori-like level-wise generation of the combinations of frequent itemsets. This dramatically reduces the size of conditional-pattern base generated at the subsequent level of search as well as the size of its corresponding conditional FP-tree. Moreover, it transforms the problem of finding long frequent patterns to looking for shorter ones and then concatenating the suffix. It employs the least frequent items as suffix, which offers good selectivity. All these techniques contribute to substantial reduction of search costs. A performance study has been conducted to compare the performance of FP-growth with two representative frequent-pattern mining methods, Apriori (Agrawal and Srikant, 1994) and TreeProjection (Agarwal et al., 2001). Our study shows that FP-growth is about an order of magnitude faster than Apriori, especially when the data set is dense (containing many patterns) and/or when the frequent patterns are long; also, FP-growth outperforms the TreeProjection algorithm. Moreover, our FP-tree-based mining method has been implemented in the DBMiner system and tested in large transaction databases in industrial applications. Although FP-growth was first proposed briefly in Han et al. (2000), this paper makes additional progress as follows. – The properties of FP-tree are thoroughly studied. Also, we point out the fact that, although it is often compact, FP-tree may not always be minimal. – Some optimizations are proposed to speed up FP-growth, for example, a technique to handle single path FP-tree has been further developed for performance improvements. – A database projection method has been developed in Section 4 to cope with the situation when an FP-tree cannot be held in main memory—the case that may happen in a very large database. – Extensive experimental results have been reported. We examine the size of FP-tree as well as the turning point of FP-growth on data projection to building FP-tree. We also test the fully integrated FP-growth method on large datasets which cannot fit in main memory.

2.9 J. Vaidya and C. Clifton, “Privacy preserving association rule mining in vertically partitioned data

Data mining technology has emerged as a means for identifying patterns and trends from large quantities of data. Mining encompasses various algorithms such as clustering, classification, association rule mining and sequence detection. Traditionally, all these algorithms have been developed within a centralized model, with all data being gathered into a central site, and algorithms being run against that data. Privacy concerns can prevent this approach – there may not be a central site with authority to see all the data. We present a privacy preserving algorithm to mine association rules from vertically partitioned data. By vertically partitioned, we mean that each site contains some elements of a transaction. Using the traditional “market basket” example, one site may contain grocery purchases, while another has clothing purchases. Using a key such as credit card number and date, we can join these to identify relationships between purchases of clothing and groceries. However, this discloses the individual purchases at each site, possibly violating consumer privacy agreements. There are more realistic examples. In the sub-assembly manufacturing process, different manufacturers provide components of the finished product. Cars incorporate several subcomponents; tires, electrical equipment, etc.; made by independent producers. Again, we have proprietary data collected by several parties, with a single key joining all the data sets, where mining would help detect/predict malfunctions. The recent trouble between Ford Motor and Firestone Tire provide a real-life example. Ford Explorers with Firestone tires from a specific factory had tread separation problems in certain situations, resulting in 800 injuries. Since the tires did not have problems on other vehicles, and other tires on Ford Explorers did not pose a problem, neither side felt responsible. The delay in identifying the real problem led to a public relations nightmare and the eventual replacement of 14.1 million tires[16]. Many of these were probably fine – Ford Explorers accounted for only 6.5 million of the replaced tires [11]. Both manufacturers had their own data – early generation of association rules based on all of the data may have enabled Ford and Firestone to resolve the safety problem before it became a public relations nightmare. Informally, the problem is to mine association rules across two databases, where the columns in the table are at different sites, splitting each row. One database is designated the primary, and is the initiator of the protocol. The other database is the responder. There is a join key present in both databases. The remaining attributes are present in one database or the other, but not both. The goal is to find association rules involving attributes other than the join key. We must also lay out the privacy constraints. Ideally we would achieve complete zero knowledge, but for a practical solution controlled information disclosure may be acceptable. Finally, we need to quantify the accuracy and the efficiency of the algorithm, in view of the security restrictions.

2.10 PRIVACY-PRESERVING DISTRIBUTED MINING OF ASSOCIATION RULES ON HORIZONTALLY PARTITIONED DATA

Data mining can extract important knowledge from large data collections¿but sometimes these collections are split among various parties. Privacy concerns may prevent the parties from directly sharing the data and some types of information about the data.

CHAPTER – III

RESEARCH METHODOLOGY

3.1. INTRODUCTION

For actual testing & getting the result by implementing new methodologies in data mining, the researcher gone through all the details about the consumer behavior & he experiment it by choosing a organization a mall or super market as a sample for his study. Researcher collect all the live data day wise, month wise i.e. transaction of each customer. After collecting the data researcher searches various methodologies which go through the methodology for finding the answer. We choose the suitable technique, formula, algorithms, methods for the customer data base. After collecting the data, researcher select the suitable method from various alternatives. He select association rule for checking the association between the various products which are bought by the customer. He implement the market Basket Analysis for this database

3.2. WEB MINING

The information space known as Web is a collection of resources (Web resources) residing on the Internet, that can be accessed using HTTP and protocols that derive from it. A resource "can be anything that has identity. Familiar examples include an electronic document, an image, a service (e.g., "today's weather report for Los Angeles"), as well as a collection of other resources. Not all resources are network "retrievable"; e.g., human beings, corporations, and bound books in a library can also be considered resources". The most important concept regarding the Web is of course the resource that a server makes available to clients spread everywhere on the Internet, without any resource, the whole system won't have any sense. When a resource is accessed by a client at a specific time and space, we talk of resource manifestation. The general definition for client is "the role adopted by an application when it is retrieving and/or rendering resources or resource manifestations", whereas the specific one for the Web defines the Web client as an application "capable of accessing Web resources by issuing requests and render responses containing Web resource manifestations" On the other hand, the server is "the role adopted by an application when it is supplying resources or resource manifestations" to the requesting client.

3.3. ASSOCIATION RULE MINING

Many machine learning algorithms that are used for data processing and information science work with numeric information. and plenty of algorithms tend to be terribly mathematical (such as Support Vector Machines, that we have a tendency to antecedently discussed). But, association rule mining is ideal for categorical (non-numeric) information and it involves very little over straightforward counting! That's the sort of algorithm that MapReduce is absolutely smart at, and it also can cause some extremely fascinating discoveries.

Association rule mining is primarily centered on finding frequent co-occurring associations among a set of things. It's typically brought up as "Market Basket Analysis", since that was the initial application space of association mining. The goal is to seek out associations of things that occur along a lot of usually than you'd expect from a sampling of all prospects. The classic example of |this can be the celebrated brewage and Diapers association that's often mentioned in data mining books. The story goes like this: men who head to the shop to shop for diapers will tend to shop for brewage at an equivalent time. Allow us to illustrate this with an easy example. Suppose that a store's retail transactions database includes the subsequent information:

There are 600,000 transactions in total.

- 7,500 transactions contain diapers (1.25 percent)
- 60,000 transactions contain beer (10 percent)
- 6,000 transactions contain each diapers and beer (1.0 percent)

If there was no association between beer and diapers (i.e., they are statistically independent), then we have a tendency to expect solely 100% of diaper purchasers to conjointly obtain brewage (since 100% of all customers obtain beer). However, we tend to discover that 80th (=6000/7500) of diaper purchasers conjointly obtain brewage. this is often an element of eight increase over what was expected – that's known as raise, that is that the quantitative relation of the ascertained frequency of co-occurrence to the expected frequency. This was firm just by investigation the transactions within the info. So, during this case, the association rule would state that diaper purchasers will obtain brewage with a raise issue of 8. In statistics, raise is just estimated by the quantitative relation of the chance of two things x and y, divided by the merchandise of their individual probabilities: raise = P(x,y)/[P(x)P(y)]. If the two things ar statistically freelance, then P(x,y)=P(x)P(y), reminiscent of raise = one therein case. Note that anti-correlation yields raise values but one, that is additionally a noteworthy discovery – similar to reciprocally exclusive things that seldom co-occur along.

The on top of straightforward example was created up, and it's terribly rare in planet cases to possess raise factors as high as eight. But, there was a case wherever it did happen. That case was discovered by Walmart in 2004 once a series of hurricanes crossed the state of FL. when the primary cyclone, there have been many a lot of hurricanes seen within the Atlantic Ocean heading toward FL, and then Walmart mined their large retail dealings info to examine what their customers extremely wished to shop for before the arrival of a cyclone. They found one specific item that exaggerated in sales by an element of seven over traditional searching days. That was an enormous raise issue for a real-world case. That one item wasn't drinking water, or batteries, or beer, or flashlights, or generators, or any of the standard things that we would imagine. The item was strawberry pop tarts! One may imagine scores of reasons why this was the foremost desired product before the arrival of a cyclone – pop tarts don't need refrigeration, they are doing not ought to be burned, they are available in severally wrapped parts, they need an extended period, they're a dish, they're a food, youngsters love them, and that we love them. Despite these “obvious” reasons, it absolutely was a still an enormous surprise! and then Walmart furnished their stores with plenty of strawberry pop tarts before succeeding hurricanes, and that they oversubscribed them out. that's a win-win: Walmart wins by creating the sell, and customers win by obtaining the merchandise that they most need.

Another example of association mining was provided to Pine Tree State by a colleague of mine at Mason University. he's a academician of geoinformation systems and natural science. He used this algorithmic program to look at the characteristics of hurricanes (internal wind speed, gas pressure within the eye of the cyclone, wind shear, rain amounts, direction and propagation speed of the cyclone, etc.), and he found a powerful association between the ultimate strength (category) of the cyclone and also the values of these totally different characteristics. He was able to predict cyclone intensification and its final strength a lot of accurately with association mining than the quality cyclone model employed by the national cyclone center. That was an incredible application of associate degree algorithmic program that was initially developed for place of business dealings mining.

3.3.1. History

The thought of association rules was popularized significantly because of the 1993 article of Agrawal et al., that has noninheritable over eighteen,000 citations in line with Google Scholar, as of August 2015, and is therefore one in every of the foremost cited papers within the data mining field. However, it's doable that what's currently known as "association rules" is comparable to what seems within the 1966 paper on GUHA, a general data mining methodology developed by Petr Hájek et al. An early (circa 1989) use of minimum support and confidence to seek out all association rules is that the Feature primarily based Modeling framework, that found all rules with (X)} }(X) and (X\Rightarrow Y)} (X\Rightarrow Y)} larger than user outlined constraints.

Alternative measures of interest

In addition to confidence, different measures of interest for rules are projected. Some widespread measures are:

- All-confidence
- Collective strength
- Conviction
- Leverage
- Lift (originally known as interest)

Several a lots of measures are given and compared by Tan et al and by Hahsler. Longing for techniques which will model what the user has illustrious (and victimization these models as interest measures) is presently an energetic analysis trend beneath the name of "Subjective interest."

Statistically sound associations One limitation of the quality approach to discovering associations is that by looking out large numbers of doable associations to look for collections of things that appear to be associated, there's an oversized risk of finding several spurious associations. These are collections of things that go with surprising frequency within the information; however solely do thus out of the blue. for instance, suppose we have a tendency to are considering a set of ten,000 things and looking out for rules containing two things within the left-hand-side and one item within the right-hand-side. There are or so 1,000,000,000,000 such rules. If we tend to apply a statistical check for independence with a significance level of zero.05 it suggests that there's solely a 5-hitter likelihood of exceptive a rule if there's no association. If we have a tendency to assume there aren't any associations, we should always withal expect to seek out 50,000,000,000 rules. Statistically sound association discovery controls this risk, in most cases reducing the danger of finding any spurious associations to user-specified significance levels.

3.4 Frequent Pattern Mining

Frequent pattern mining has been the main focus of nice interest among data mining researchers and practitioners. It's these days wide accepted to be one among the key issues within the data mining fields. Frequent pattern mining is to search out the item sets that seem ofttimes from much knowledge; an item set is any set of the set of all things. Frequent pattern mining is that the discovery of relationships or correlations between things in a very dataset. Within the case of information streams, one might need to search out the frequent item sets either over a window or the whole knowledge stream. Frequent pattern drawback: The frequent pattern mining problem was 1st introduced as mining association rules between sets of things. Let I = be a group of things. An itemset X I could be a set of things. We have a tendency to write itemsets as X = ij1 … ijn, i.e. Omitting set brackets. Notably, a thingset with l items is named an l-itemset. A transaction T = (tid, X) could be a tuple wherever tid could be a transaction-id and X is an itemset. A transaction T = (tid, X) is alleged to contain itemset Y if Y X. A transaction info db could be a set of transactions. The support of an itemset X in transaction info db, denoted as supdb(X) or sup(X), is that the range of transactions in db containing X, i.e., sup(X) = | ((tid, Y) Є DB) Λ (X Y)| drawback statement: Given a user-specified support threshold min sup, X is named a frequent itemset or frequent pattern if sup(X) ≥ minutes sup.

The matter of mining frequent itemsets is to search out the whole set of frequent itemsets in a very dealings info db with regard to a given support threshold minutes sup. To check frequent pattern mining in knowledge streams, we have a tendency to 1st examine identical drawback in very dealings info. To justify whether or not one item i1 is frequent in a very dealings info D, one simply got to scan the info once to count the quantity of transactions that i1 seems. One will count each single item i1in one scan of D. However, it's too pricey to count each doable combination of single things (i.e., itemset I of any length) in D as a result of there ar a large range of such mixtures. An economical different planned is that the Apriori rule to count solely those itemsets whose each correct set is frequent. That is, at the k'th scan of D, derive its frequent itemset of length k (where k &gt;= 1), then derive the set of length (k+1) candidate itemset (i.e., whose each length k set is frequent) for consecutive scan.

Frequent pattern mining could be a rather broad space of analysis, and it relates to a good style of topics a minimum of from an application specific-perspective. Loosely, the analysis within the space falls in one among four completely different categories: Technique-centered: This space relates to the determination of a lot of economical algorithms for frequent pattern mining. A good style of algorithms is planned during this context that use completely different enumeration tree exploration methods, and completely different knowledge illustration ways. Additionally, various variations like the determination of compressed patterns of nice interest to researchers in data mining. quantifiability problems: The quantifiability issues in frequent pattern mining are terribly vital. Once the info arrives within the type of a stream, multi-pass ways will not be used. Once the info is distributed or terribly massive, then parallel or big-data frameworks should be used. These situations necessitate differing types of algorithms.

Advanced knowledge varieties: various variations of frequent pattern mining are planned for advanced knowledge types. These variations are used in a very large choice of tasks. Additionally, completely different knowledge domains like graph knowledge, tree structured knowledge, and streaming knowledge typically need specialised algorithms for frequent pattern mining. Problems with powerfulness of the patterns also are quite relevant during this context. Applications: Frequent pattern mining have various applications to alternative major data mining issues, net applications, code bug analysis, and chemical and biological applications. A major quantity of analysis has been dedicated to applications as a result of these are notably necessary within the context of frequent pattern mining.

3.4.1. Frequent Itemset Generation

Here the target is to search out all the itemsets that satisfy the minsup threshold. These itemsets are known as frequent itemsets. The procedure necessities for frequent itemset generation are quite dearly-won. A lattice structure is employed to enumerate the list of seventy two all doable itemsets. In general, a knowledge set that contains k things will doubtless generate up to 2k-1 frequent itemsets, excluding the set. As a result of k may be terribly massive in several sensible applications, the search house of itemsets that require to be explored is exponentially massive.

3.5 PROBLEM DEFINITION

A number of privacy-preserving mining solutions have been proposed in recent times. In their settings, there are multiple data owners wishing to learn association rules or frequent itemsets from their joint data. However, the data owners are not willing to send their raw data to a central site due to privacy concerns. If each data owner has one or more rows (i.e. transactions). In the joint database, we say that the database is horizontally partitioned. If each data owner has one or more columns in the joint database, the database is considered vertically partitioned. This paper focuses on vertically partitioned databases; such databases are useful for market basket analysis. A transaction of the database contains the products that a customer had bought from one or more of the participating businesses, and attributes such as the customer credit card number and date of purchase are used as TIDs. Therefore, each of the businesses (i.e. data owners) will own some transaction partitions in the joint database. However, these businesses may not wish to disclose such data, which include trade secrets (e.g. there may be other competing businesses sharing the same joint database) and customer privacy (e.g. due to regulations in existing privacy regime). Therefore, a privacy preserving mining solution must be applied. Other use cases can also be found in areas such as automotive safety and national security.

OBJECTIVES

The goals of our proposed privacy-preserving association rule mining and frequent itemset mining for vertically partitioned databases are as follow:

Privacy: Data owners should learn as little information about databases belonging to other data owners as possible. More specifically, a data owner’s raw transaction details should not be disclosed, and the supports should be concealed to avoid leakage of information about the raw data. Similarly, exact confidences should be concealed as they could be used to infer some information about the raw data. The proposed solutions should also protect the mining results from the cloud.

Efficiency: Privacy-preserving measures usually result in decreased performance of data mining, and therefore, any trade-off has to be realistic. In our context, the data mining latency should be acceptable compared with the latencies of non-privacy-preserving data mining algorithms.

3.6 EXISTING SYSTEM

In this system, studied the problem of outsourcing the association rule mining task within a corporate privacy-preserving framework. A substantial body of work has been done on privacy-preserving data mining in a variety of contexts. A common characteristic of most of the previously studied frameworks is that the patterns mined from the data (which may be distorted, encrypted, anonymized, or otherwise transformed) are intended to be shared with parties other than the data owner. The key distinction between such bodies of work and our problem is that, in the latter, both the underlying data and the mined results are not intended for sharing and must remain private to the data owner.

3.6.1 DISADVANTAGES

- It is prohibitively expensive to achieve perfect secrecy of outsourced frequent itemset mining
- They do not offer any theoretical analysis of anonymity of item sets.

3.7 PROPSOED SYSTEM

In this system, proposes privacy-preserving mining solutions for high privacy requirements. This paper proposes an efficient homomorphic encryption scheme and a secure outsourced comparison scheme. To avoid the disclosure of supports/confidences, we design an efficient homomorphic encryption scheme to facilitate secure outsourced computation of supports/ confidences, as well as a secure outsourced comparison scheme for comparing supports/confidences with thresholds. The proposed (symmetric homomorphic) encryption scheme is tailored for the proposed comparison. The scheme only requires modular additions and multiplications, and is more efficient than the homomorphic encryption schemes used in other association rule mining and frequent itemset mining solutions.

3.7.1 ADVANTAGES

- They can potentially be adopted in a wide range of secure computation applications
- It is more efficient than the homomorphic encryption schemes used in other association rule mining and frequent itemset mining solutions.

3.8. MODULE DESIGN

User Management:

This module deals with user details of this application. The main interactive users on this application are administrator and Customers. The administrator is who manage the entire shopping cart application and the customer is a person who orders products through this application. The Registered customer only can order products through this application

Product Management:

In this module enables the administrator to add new product or edit/delete existing product on this application. All the Product are added by the specified category like Power, Telecom, Network, Modems etc.,

Order Management:

In this module describes the customer Order details. The customer can search any product details on this website and can make order through order entry module. The order details are stored in order database and searching details will be stored in searching log file. The order details will be used there after to predict missed items details of the customer.

Delivery Management:

This module manages delivery details. After Payment to the company the product will be delivered to the customers. The company deals with prestigious company, so their delivery management is monitored by the management people. Delivery status will be remained to the ordered customers via e-mail.

Market Basket Prediction Module:

This is the important module on this application; it enables the administrator to make analysis of customer preferred items from customer orders. When the admin select a customer, it displays all the order details and most liked product details of the selected customer. Preference table analyzes the customer’s choice of interest and let the admin to know the customer’s preference either Brand (company) or Features adopted within the product.

Customer Registration:

Before ordering products the customer has to register their details in online. Every registered customer has privileges to order products. The customer by himself can have their login details after completing registration.

Comparative Chart:

The admin uploaded product details will be displayed in chart format. It includes the product name, the company’s selling the product (i.e. their partners), the features offered by each brands. And the customers have the option to select multiple products of their choice from the given chart.

Search and Order Product:

The customer will search for their interest of products to order. The categories arranged by the admin of this project, easies the work of customer to find his product. As well as by specifying the type of product the customer will get the product details list. From the results they may order to the company. Later on the prediction analysis system implemented, the customer will get their missing orders from their logs.

Comments/Feedback:

The customer has the facility to send their comments and feedback to the administrator. This will be checked and replied by the admin.

Product Ranking:

Additional feature added with this project is ranking the company's selling product. Based on the sales the product selling estimation will be monitored and logged in database. This will be analyzed and shows Top 5 selling products of the company. This prediction will increase the sales of the company and knows the user crowd willing to buy that product.

CHAPTER – 4

RESULTS AND DISCUSSIONS

4.1. REVIEW CHART FOR CUSTOMER USING PRICE

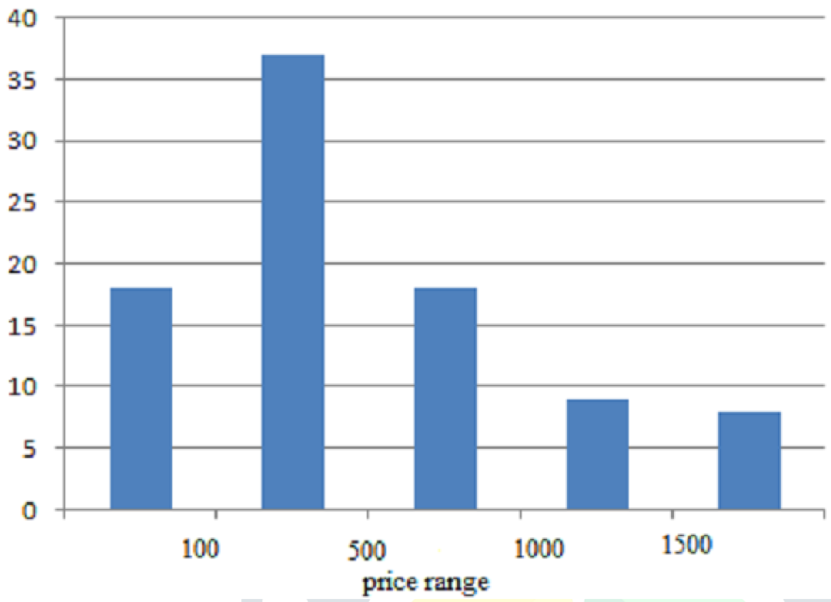

Fig 4.1. The output shown in a graphical form. X-axis represents the price range and Y-axis represents the number of customers.

Researcher selects Market Basket Analysis for his data analysis because Market Basket analysis is a tool of knowledge discovery about co-occurrence of nominal or categorical items. Market Basket Transaction or market Basket Analysis is a data mining technique to derive association between data sets. Researcher has categorical data of transaction records as input to the analysis and the output of the analysis is association rules as a new knowledge directly from data.

A typical process, the researcher finds from the association rule mining is market basket analysis, This process analyzes customer buying habits by finding association between different items that customers places their "shopping basket".

The discovery of such association can help retailers develop marking strategies by gaining insight into which items are frequently purchased together by customers. For instance; if customers are buying milk, now likely are they to also buy bread on the same trip to the supermarket? Such information can lead to increased sales by helping retailers to selective marketing and plan their shelf space. For example placing milk and bread within close proximity may further encourage the sale of these items together within single visit to the store.

Initially the study applied the Association rule based mining algorithm for 1-itemset, for which I obtained the following output. The occurrence is the number of appearances of the different products in the market baskets. Again, we have only shown a snippet of the output as the ouput is very lengthy. `Then the study moved on to a 2-itemset approach and devised the rules for 2 – itemsets based on the same datasets. The output had hundreds of rules which then had to be pruned down on the basis of the "coverage values".

- In terms of predictions, the clusters obtained can show the different segments of customers and the more populated segments can be targeted specifically.
- The customer purchase patterns approach, using the association rules mining technique, is an effective way of extracting the rules from the raw data and inferring the buying patterns among them.
- The implementation shows that increasing the "coverage" values results in better pruning of rules, and a more trustworthy rule set.
- Also, the increase from a 1-item to a 2-itemset and then onto a 3-itemset, results in a 10-folds increase in the computation times of the algorithms.
- The changing of loop-nesting from a pre-conditioned one to the post-condition one does reduce the execution time a bit but not very significantly.

CHAPTER – 5

CONCLUSION AND FUTURE ENHANCEMENTS

5.1. CONCLUSION

Data Mining System is useful to study buying behavior of the customers in retail departmental stores. With this study researcher has concluded that there are certain buying habits of the customers. And according to this buying habits of customer, management may update their system of providing various types of services to their customers to delight the customers and to retain the customer with same business house. 2) The data mining system is useful to Business house to find out the association of the customers with different products. And how customers are shifting from one brand to another brand of product to satisfy their need because their earlier buying habits are properly studied by the Data mining System.

5.2. FUTURE ENHANCEMENTS

The work done so far leaves ample amount of space for future improvements and comparisons with other algorithms. The Association rule algorithm implemented here takes up a lot of execution time. So, the optimization of the algorithm can be done to ensure a better performing algorithm. The Association Rule mining algorithm implemented but the running time is not good. Either, these values can be taken into account 1 at a time or, all three at a time using weight factors for the three parameters. Of course, here a simple database was used, so there isn't much space for taking the frequency values into account. But it can be extended to a larger and more comprehensive database to analyze the aforementioned values.

CHAPTER – 6

REFERENCES

[1] Yoseph Linde, Andres Buzo, Robert M. Gray : An Algorithm for Vector Quantizer Design, IEEE Transactions on communications, vol. com-28, no. 1, (january 1980), pp. 84-86.

[2] Danuta Zakrzewska, Jan Murlewski : Clustering Algorithms for Bank Customer Segmentation, 5th International Conference on Intelligent Systems Design and Applications, (2005),pp 1-2.

[3] Abdullah Al-Mudimigh, Farrukh Saleem, Zahid Ullah Department of Information System: Efficient implementation of data mining: improve customer's behavior, 2009 IEEE ,(2009),pp.7-10.

[4] Sung Ho Ha , Sang Chan Park, Sung Min Bae : Customer's time-variant purchase behavior and corresponding marketing strategies: an online retailer's case, Computers & Industrial Engineering 43 (2002) 801–820, (2002),pp.801-806.

[5] Euiho Suh, Seungjae Lim, Hyunseok Hwang, Suyeon Kim : A prediction model for the purchase probability of anonymous customers to support real time web marketing: a case study, Expert Systems with Applications 27 ,(2004), pp. 245-250.

[6] Mu-Chen Chen , Hsu-Hwa Chang, Ai-Lun Chiu : Mining changes in customer behavior in retail marketing, Expert Systems with Applications 28 ,(2005), pp. 773-776.

[7] Sriram Thirumalai, Kingshuk K. Sinha : Customer satisfaction with order fulfillment in retail supply chains: implications of product type in electronic B2C transactions, Journal of Operations Management 23 ,(2005), pp. 291-296.

[8] Ian H. Witten & Eibe Frank : Data Mining : Practical machime learning tools and techniques, San Francisco, Morgan Kaufmann publishers, (2005), pp. 112-118,136-139.

[9] Ajay Agrawal, Joshua Gans, and Avi Goldfarb. 2017. How AI will change strategy: A thought experiment. Harvard Business Review (2017).

[10] Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast algorithms for mining association rules. In International Conference on Very Large Data Bases.

[11] Tsan-Ming Choi, Chi-Leung Hui, Na Liu, Sau-Fun Ng, and Yong Yu. 2014. Fast fashion sales forecasting with limited data and time. Decision Support Systems 59 (2014), 84–92.

[12] Tsan-Ming Choi, Yong Yu, and Kin-Fan Au. 2011. A hybrid SARIMA wavelet transform method for sales forecasting. Decision Support Systems 51 (2011), 130–140.

[13] Chad Cumby, Andrew Fano, Rayid Ghani, and Marko Krema. 2004. Predict-ing customer shopping lists from point-of-sale purchase data. In ACM SIGKDDInternational Conference on Knowledge Discovery and Data Mining (KDD).

[14] Riccardo Guidotti, Giulio Rossetti, Luca Pappalardo, Fosca Giannotti,and DinoPedreschi. 2017. Market basket prediction using user-centric temporal annotated
recurring sequences. In IEEE International Conference on Data Engineering.

[15] Riccardo Guidotti, Giulio Rossetti, Luca Pappalardo, Fosca Giannotti, and Dino Pedreschi. 2018. Personalized market basket prediction with temporal annotated recurring sequences. IEEE Transactions on Knowledge and Data Engineering(2018).

[16] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In International Conference on World Wide Web (WWW).

[17] Balázs Hidasi and Alexandros Karatzoglou. 2018. Recurrent neural networks with top-k gains for session-based recommendations. In International Conference on Information and Knowledge Management.

[18] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based recommendations with recurrent neural networks. In Inter-national Conference on Learning Representations (ICLR).

[19] Bruno J. D. Jacobs, Bas Donkers, and Dennis Fok. 2016. Model-based purchase predictions for large assortments. Marketing Science 35, 3 (2016), 389–404.

[20] Rohit J. Kate. 2016. Using dynamic time warping distances as features for im-proved time series classification. Data Mining and Knowledge Discovery 30, 2 (2016), 283–312.

[21] Mathias Kraus, Stefan Feuerriegel, and Asil Oztekin. 2018. Deep learning in business analytics and operations research: Models, applications and managerial implications. arXiv preprint arXiv:1806.10897 (2018).

[22] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In International Conference on Machine Learn-ing (ICML).

[23] Moshe Lichman and Padhraic Smyth. 2018. Prediction of sparse user-item con-sumption rates with zero-inflated poisson regression. In International Conference on World Wide Web (WWW).

[24] Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing 7, 1 (2003), 76–80.

[25] A.L.D. Loureiro, V. L. Miguéis, and Lucas F.M. da Silva. 2018. Exploring the use of deep neural networks for sales forecasting in fashion retail. Decision Support Systems 114 (2018), 81–93.

[26] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems (NeurIPS).

[27] Ofir Pele and Michael Werman. 2009. Fast and robust earth mover’s distances. In International Conference on Computer Vision (ICCV).

[28] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In Conference on Uncertainty in Artificial Intelligence (UAI).

[29] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factoriz-ing personalized markov chains for next-basket recommendation. In International Conference on World Wide Web (WWW).

[30] Yasushi Sakurai, Christos Faloutsos, and Masashi Yamamuro. 2017. Stream monitoring under the time warping distance. In IEEE International Conference on Data Engineering.

[31] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In International Conference on World Wide Web (WWW).

[32] Nhat-Quang Tran, Ba-Ngu Vo, Dinh Phung, and Ba-Tuong Vo. 2016. Clustering for point pattern data. In International Conference on Pattern Recognition.

[33] Ramachandran Varatharajan, Gunasekaran Manogaran, Malarvizhi Kumar Priyan, and Revathi Sundarasekar. 2018. Wearable sensor devices for early detection of Alzheimer disease using dynamic time warping algorithm. Cluster Computing 21, 1 (2018), 681–690.

[34] Pengfei Wang, Jiafeng Guo, Yanyan Lan, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2015. Learning hierarchical representation model for next basket recom-mendation. In International Conference on Research and Development in Informa-tion Retrieval.

[35] Jason Weston, Hector Yee, and Ron J. Weiss. 2013. Learning to rank recommenda-tions with the k-order statistic loss. In ACM Conference on Recommender Systems
(RecSys).

[36] Peter R. Winters. 1960. Forecasting sales by exponentially weighted moving averages. Management Science 6, 3 (1960), 324–342.