Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 136 (2018) 246–254

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Young Scientist Conference on Computational Science

Market basket analysis of heterogeneous data sources for recommendation system improvement

Kutuzova Tatiana[a*], Melnik Mikhail[a]

ITMO University, 49 Kronverksky Pr., Saint-Petersburg 197101, Russia

Abstract

Information systems that process a large amount of data become an integral part of our lives. Development of online markets and market technologies lead to the need for retailers to analyze customers' behaviour. The result of the effective analysis may improve both supplier's profitability, quality of service and customer satisfaction that attracts increased interest for research. One of retailing data analytics applications is the construction of recommendation system. Increase the quality of the recommendation system is possible when analyzing a larger amount of data, which can be obtained from external heterogeneous sources. Examples of sources for data integration can be data from online and offline markets inside one company or data from partner companies. Within one market area, a range of offered products may be similar, while the characteristics or associative rules formed for them may differ. Therefore, for the correct integration of external data sources into the existing recommendation system, it is required to analyze the structure and content of additional data sources to use only beneficial parts of that data. In this work, we propose a study on the integration of heterogeneous data sources from a grocery supermarket based on Market Basket Analysis methods.

Keywords: market basket analysis; recommendation system; association rules; data integration

1. Introduction

Today, information systems that process a large amount of data become an integral part of our lives. Development of online markets and market technologies lead to the need for retailers to analyze customers' behaviour. The result of

* Corresponding author. Tel.: +7-921-571-7693.
E-mail address: kutuzova.tanya@mail.ru

10.1016/j.procs.2018.08.263

the effective analysis may improve both supplier's profitability, quality of service and customer satisfaction that attracts increased interest for research.

Market Basket Analysis or MBA is a field of modelling techniques based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items [1]. MBA includes determination and prediction customer's behaviour based on expenditure pattern of previous clients. MBA is applied not only in retail but also in a great number of other fields. There are studies which point to MBA and contribute to increasing incomes in hotels management by offering more attractive additional services for new and regular customers. MBA based on multidimensional logit-model was used to conduct a study [2]. Another challenge is to make a choice of purchasing, sailing or ownership of stocks in an equity market. Data mining techniques ensure high precision of prediction of stock price movement [3]. Using of MBA is not merely limited to using in fields which depend on people and their behaviour, it also uses in biology, chemistry and ecology [4], [5]. In the study [6] using MBA for improving methods of arranging products on store shelves was identified. Analysis of the most frequent customers' transactions was performed.

Association rules (AR) is a widespread technique in different fields. AR is used to identify relations in big datasets of transactions. For instance, AR are used in cross-marketing and in decisions making in business processes determining of significant relations [7]. Moreover, AR are used to the determinate a risk factor for brain cancer[8].

A finding of significant item sets in a big dataset is one of the main researches in data mining. MBA encompasses development and analysis association rules which look like 'Men who buy diapers for their kids, are also more likely to have beer in their carts'[9]. For AR, it is necessary to analyze co-purchased and often occurred together items from set of transactions.

Objects from one market area provide similar assortments of goods or services with different characteristics. For example, an assortment of supermarkets can be varied in diversity or quality of products. Thus, customers' behaviour can also differ, that leads to a uniqueness in a transaction database, which is required for constructing association rules by generating frequent sets of elements and generating rules. There are several algorithms, including the most common "Apriori algorithm".

"Apriori Algorithm" is one of the data mining algorithms that used for MBA and mining potential AR. This algorithm was proposed by Agrawal and Srikant in 1994. It is one of the most popular algorithms for AR mining in big datasets and uses the downward closure property [10].

Recommendation system (RS) is an AI algorithm that filter information about customers' behaviour and suggests additional products to them. It is based on a variety of factors such as past purchases, demographic info, their search history and etc. Implementation of RS has three main approaches: collaborative filtering, content-based filtering and their hybrid recommendation systems[11].

Collaborative filtering is based on past customers' behaviour and it collects and analyzes a large amount of information about users (purchases, activities, preferences etc.) and then predicts what users will prefer based on similarity with other users. Some collaborative filtering algorithms are surveyed and compared in [12]. Moreover, this approach is used in Amazon nowadays[13].

Content-based filtering is based on similar properties of products and user's profiles of preferences. In [14] content-based filtering is used for recommendation of news topics.

The last hybrid recommendation systems combine principles of two previous approaches and one of an examples of that approach is presented in [15] for multi-criteria collaborative filtering.

Quite often the arrangement of goods on shelves of a shopping room and online recommendations of stores do not correlate with each other, that leads to inefficient use of the store's resources. In retail, the analysis of purchases (transactions) is a crucial part of understanding the customer's behaviour. The data on purchases allow shops to adjust promotions, store individual preferences and improve the quality of service [6]. Therefore, in this study, we are trying to improve the quality of built recommendation system by using external heterogeneous data sources.

2. Methodology

2.1. Association rules

MBA is a bunch of algorithms that detect patterns of purchases by mining of associations or joint events from transaction's database. Our recommendation system consists of two objects that used to provide recommendations.

The formal definition of mining AR is described in [16]. Let I is a set of m items $I = I_1, I_2, \dots, I_m$, database D with various of transactions T, transaction T_s consist of some I, $T_s \subseteq I$,. AR has the form $X \Rightarrow Y$, where $X, Y \subset I$ is elements which called itemsets and $X \bigcap Y = \emptyset$, X is antecedent and Y is consequent. The rule implies that X contains Y.

The strength of rules is calculating for this and consists of three measures [16]: support, confidence and lift.

Support – fractions of transactions, which consist of all items in the itemset. The higher the value of support, the more often itemset occurs. AR with high support is preferable to apply to more number of future transactions. Value of support can be calculated by the next equation:

$$S(X) = \frac{count(X)}{N} \tag{1}$$

where N is na umber of all transactions.

Confidence – a probability that items from the right side of a rule will be bought with items from the left side of the rule and calculated by:

$$C(X \Rightarrow Y) = \frac{S(X \cup Y)}{S(X)} \tag{2}$$

The last significant measure is a lift, the probability of co-occurring items in rules divided to the multiplication of supports of left and right sides. Lift sum up the strength of AR between items in the left and right sides and calculated by the formula:

$$L(X \Rightarrow Y) = \frac{S(X \cup Y)}{S(X)S(Y)} = \frac{C(X \Rightarrow Y)}{S(Y)} \tag{3}$$

The higher value of lift, the stronger connection between items.

Implementation of the Apriori algorithm has two steps [7], [17]. At first, finding all item sets, which frequently occur in transactions dataset with support more than a minimum threshold of support, which was set by a user.

Table 1. An example of frequent itemsets.

Support	Itemsets
0.1502	Whole milk
0.1079	Herbs
0.0639	Chicken
0.0614	Root vegetables
0.0610	Spread cheese

After that, confidence calculating for all possible rules based on frequent itemsets and save rules with confidence value higher than a minimum threshold which was set by the user.

Table 2. An example of association rules.

Antecedants	Consequents	Support	Confidence	Lift
Spread cheese	Whole milk	0.0102	0.1666	1.1087

Whole milk	Spread cheese	0.0102	0.0676	1.1087
Root vegetables	Herbs	0.0094	0.1535	1.4218
Herbs	Root vegetables	0.0094	0.0873	1.4218
Herbs	Salt	0.0043	0.0394	1.0302

In addition, for our recommendation system, we use the main idea of item-to-item collaborative filtering [13]. Initial recommendations are generated based on a similarity of clients, which can be measured by the cosine of the angle between two vectors:

$$sim(\vec{A},\vec{B}) = \cos(\vec{A},\vec{B}) = \frac{\vec{A}\vec{B}}{\|\vec{A}\|*\|\vec{B}\|} \quad (4)$$

By this measure, the similarity matrix can be generated, and preferable products for a client can be chosen from this matrix and used in mining AR for final recommendation.

2.2. Recommendation system

Let we have a target information system that was created for a local grocery market. The system gathers data with customers' orders and stores it in transaction database D. Supermarket offers a specific set of *n* products that customers can buy:

$$I = \{i_1, i_2, \dots, i_n\} \quad (5)$$

Therefore, transaction database $D = \{t_1, t_2, \dots t_m\}$ represents a set of all of *m* orders. Each transaction or order is a set of bought products:

$$t_j = \{i_l, i_k, \dots i_n\} \quad (6)$$

Transactions can be represented as vectors of length n with values in $\{0,1\}$, where 1 means the presence of product and 0 – absence. This representation is used for all algorithms and methods in this paper:

$$t_{j,k} = \begin{cases} 1, if\ i_k \in t_j \\ 0, otherwise \end{cases} \quad (7)$$

The built recommendation system (RS) is based on this transactions data $R(D)$. The goal of RS is to offer the most likely bought products for customer depending on his current purchases (basket). Thus, a recommendation is a function that gives a set of already chosen products and responds with a set of recommended products:

$$r(t_{input}) = t_{output} \quad (8)$$

In this work, recommendation systems based on such MBA methods as association rules and collaborative filtering. Based on a given dataset D, construction of $R(D)$ consists of following steps:

- calculate $n \times n$ similarity matrix for all products $sim(D)$ by formula (1);
- build association rules $AR(D)$.

To make a recommendation for an input vector t_{input} we select for each present product in a transaction the most similar product by using similarity matrix sim. After that, we chose rules from built AR, where items from X are in t_{input} and similar products are in Y. From all these rules, we select rule with the greater confidence as a recommendation.

2.3. Integration scheme

The aim of the integration scheme, which was proposed in this study, is to improve the quality of constructed RS by using additional data sources. These external data sources may have a completely different format and differ in content. For the mutual analysis of heterogeneous data, it is necessary to convert all the data to a suitable single form. This includes analysis of data characteristics, seeking for common groups of rules by using clusterization methods and filtering of non-conforming parts of data. Further, integrated data from all available sources are used to evaluate and improve the quality of recommendation system. The proposed scheme is presented in Fig. 1.

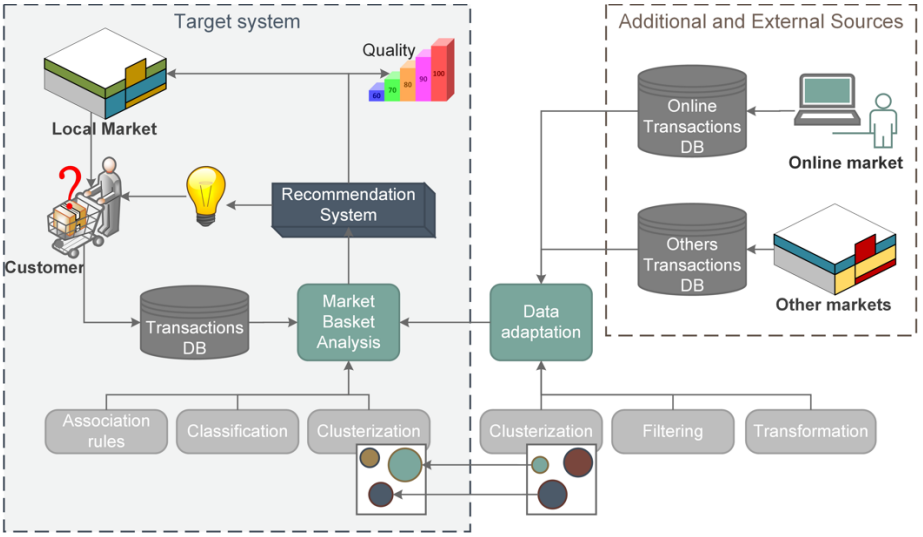

Fig. 1. Integration scheme of heterogeneous retail data sources.

Assume that we have our original dataset D_1 on which the recommendation system R was built. The dataset D_2 is an external source of data that we want to adapt and use in our RS. Sets of products I_1 and I_2 from both dataset respectively can be very different. Therefore, the first step of data integration is an adaptation of external data D_2 to the original form of D_1. If necessary, both datasets can be transformed to a new form. All transformed and filtered transactions from D_2will be stored in the new dataset D_2'.

At the next step, we perform clusterization model $C = \{c_1, c_2, \dots, c_p\}$ based on the D_1 data. For each cluster c_i we evaluate the average distance between elements inside this cluster w_i. Further, for each transaction t_j from D_2we determine the closest cluster c_i from C and calculate average distance v to points inside that cluster. To accept this transaction t_j and add it to D_2' we define criteria with a softness parameter α:

$$v < \alpha\, w_i \tag{9}$$

Therefore, we perform a filtration of all transactions from D_2 and chose only the more similar ones to original data D_1. At the final step, we construct a new recommendation system $R'(D_1 \cup D_2')$ based on the union of original and integrated datasets and then evaluate its performance.

For the evaluation, we require a prior model of customers' behaviour P. It determines the general behaviour of all possible customers. This model has the same form of the recommendation system as the others R or R'. To build a prior model, we can use data about new customers of a supermarket or use a pre-defined test sample of data.

We define two metrics to evaluate a quality of RS. The first metric (confidence distance CD) is a distance by confidence values between the same association rules $X \Rightarrow Y$ obtained from evaluating recommendation system R and prior model P. Association rules are structured as tuples in a form: (X, Y, S, C, L). Let $rule_C^R$ defines a confidence value of an association rule from R and $rule_C^P$ defines a confidence value of the same rule but from P. The number of

evaluated association rules is $|R|$. Quality of R is directly proportional to the average CD values across all association rules.

$$CD(R) = \|R, P\|_{CD} = \frac{\sum_{rule \in R} |rule_C^R - rule_C^P|}{|R|} \quad (10)$$

The second metric (recommendation conformity, RC) determines the conformity of sets with recommended products for the customers' transactions from dataset D between evaluating recommendation system R and prior model P. Function $r(t)$ corresponds to a recommendation by evaluating system R and $p(t)$ corresponds to recommendatioa n by prior model P. Quality of R is inversely proportional to the average RD values across a set of customers.

$$RC(R) = \|R, P\|_{RC} = \frac{\sum_{t \in D} \frac{|r(t) \cap p(t)|}{|r(t) \cup p(t)|}}{|D|} \quad (11)$$

3. Experimental study

3.1. Datasets

The first source of data obtained from Kaggle competition on MBA[18]. This data source includes two similar datasets of the same structure but with different sets of orders. The first one is prior dataset (K_{prior}) which contains 3214874 orders, while the second train dataset (K_{train}) contains 131209 orders. There are 49688 different products which divided by 134 aisles. These data sets were used in our experiments. K_{train} is used as a dataset to build an original recommendation system R and K_{prior} is used as a dataset to build a prior model of customers' behavior. The example of transactions from these datasets is presented in table 3.

Table 3. An example of Kaggle's datasets.

Order id	Product id	Product name	Aisle
1	49302	Bulgarian Yogurt	Yogurt
1	11109	Organic 4% Milk Fat Whole Milk Cottage Cheese	Refrigerated
36	19660	Spring Water	Water seltzer sparkling water
36	49239	Dark Chocolate Royale Shakes	Protein meal replacements

The third groceries dataset called G_{orig} was got from [19]. It has the 9834 number of orders with 169 unique products. This dataset is used as an external data source that should be adapted to construct the recommendation system. Transactions from this dataset look like in table 4.

Table 4. An example grocery (G_{orig}).

Products in orders
Whole milk, yogurt, brown bread
Tropical fruit, yogurt, coffee
Meat, citrus fruit, berries, root vegetables, whole milk, soda

3.2. Datasets unification

All datasets were reduced to a unified form to simplify future results comparison. Generalization took place in several stages. First, unique terms are defined in all datasets, such symbols as (, ; & * %) etc. were removed. Moreover, frequently encountered but not meaningful terms like 'canned', 'other', 'and' etc. were removed too. After that, all filtering words from the K_{train} and K_{prior} datasets received weights by the following formula:

$$w_i = \frac{1}{number\ of\ occurrences_i} \quad (12)$$

As a result, the most popular words got fewer weights, while less frequent words (more meaningful) received higher weights.

Names of products from all datasets were matched. Names of products with their aisles from Kaggle datasets were combined to improve future matching. As a result, all pairs of names which intersect with each other from all datasets were founded. All exceptions were removed from the list of matching. Exceptions are several rules which said which words cannot occur together. For instance, if a product has the word 'fat' product which was matched with this product cannot has words such as 'low', 'lowfat', 'low-fat'. Weights of terms were used for the reducing number of repetition; a pair of products was chosen if it has the highest weight.

After the unification of product names, the number of products was reduced to 51 items. Consequently, the number of orders were reduced too. For K_{train} dataset from 131208 to 32667, for K_{prior} dataset from 3214874 to 805894 and for G_{orig} dataset from 9835 to 8000.

3.3. Clustering

K_{train} dataset was clustered to define which orders from G_{orig} (external) dataset are more suitable to be used for RS improvement. Transactions were transformed to achieve more quality result of clustering. If initially orders were presented like a vector with 0 and 1, where 0 is the absence of a product in customers' cart and 1 is its availability, then data for clustering was transformed like where 0 is mean the same but value of a product is probability with which it can be purchased. Clusters of K_{train} were performed by K-means algorithm. Further, we select rows one by one from G_{orig} to define its proximity to K_{train}. If criterion (9) is satisfied than this transaction vector is added to additional G_{adapt} dataset, that will be used to fit recommendation system R.

3.4. Experiments

For the first part of the experiments, we construct two recommendation systems. The first one R_1 based on an original dataset G_{orig}, while the second R_2 is built on an adapted dataset G_{adapt}. Quality of both R_1 and R_2 recommendation systems were compared with prior model P by using both quality metrics – confidence distance CD and recommendation conformity RC. In this study, the prior model represents a theoretically ideal recommendation system. Results of evaluations are presented in Fig. 2.

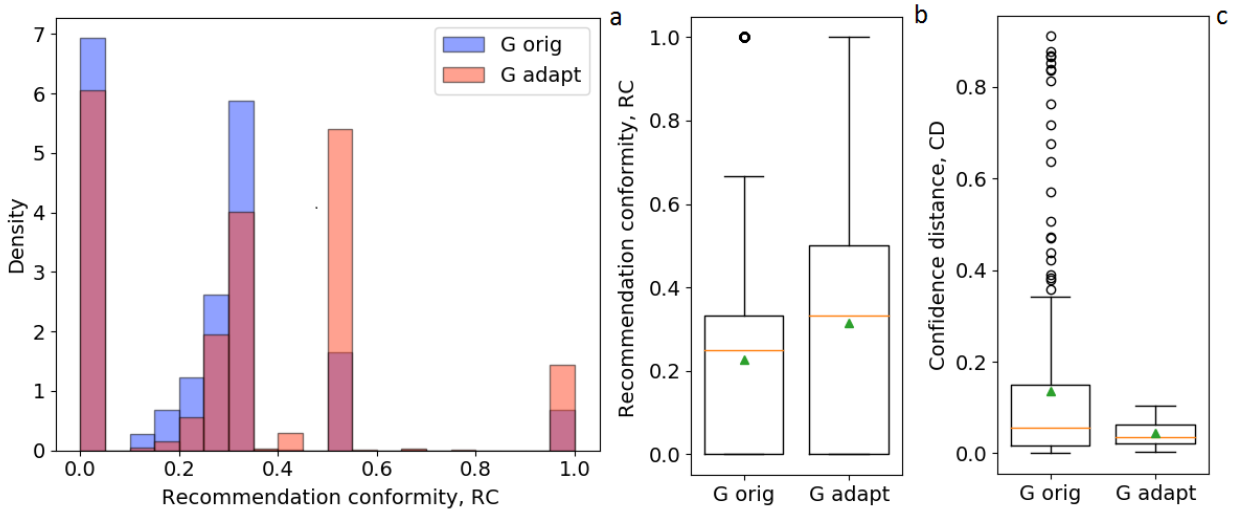

Fig. 2. Comparison of recommendation systems based on full grocery dataset and adapted grocery dataset.

The left (a) plot shows a comparison of distributions of RC values for R_1 and R_2 models across 9000 customer transactions, generated by using prior model P. The middle (b) plot shows the same results but in the form of boxplots. Adapted dataset G_{adapt} shows the better RC values, that means that quality of R_2 in compare to R_1 is greater by metric RC. The right (c) plot presents the distribution of CD values for R_1 and R_2 models. The result shows that generated on the G_{adapt} rules clearer fit the prior model than original G_{orig} data.

The second part of the experimental study performs the same steps but compares another datasets and models. Now, the first model R_2 is based on the original K_{train} data. The second R_4 recommendation model is constructed on the union of K_{train} and G_{adapt} datasets. The R_4 model represents the result of the integration of external G_{adapt} data to the initially constructed recommendation system R_3. The result of comparison of models R_3 and R_4 is presented in the same form in Fig. 3.

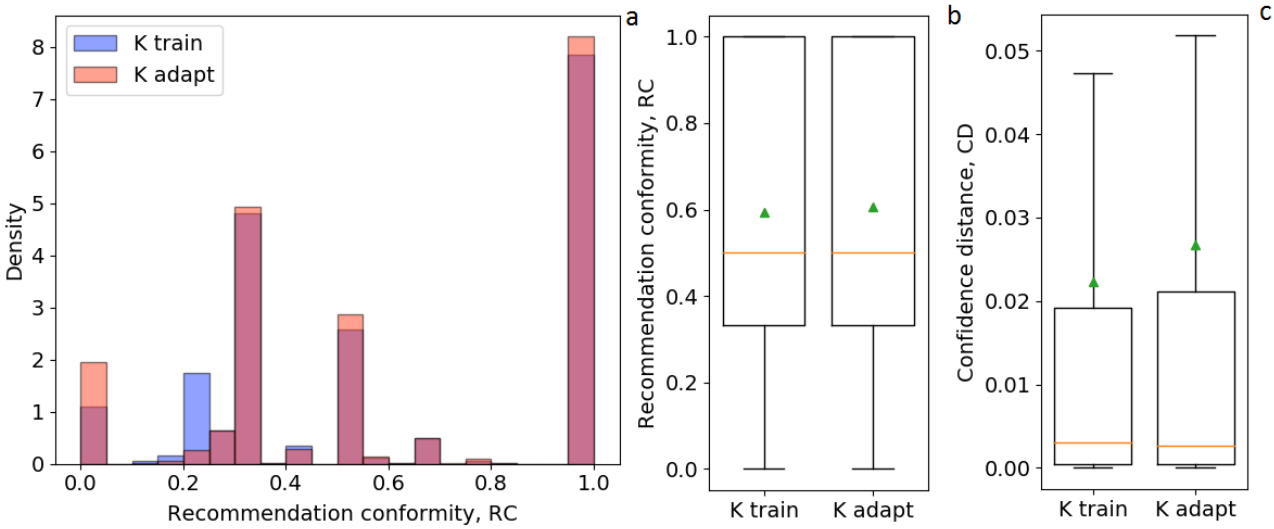

Fig. 3. Comparison of recommendation systems based on full K_{train} dataset and adapted K_{train} dataset.

Now, results show neutral performance for integrated recommendation model R_4 in compare to the original model R_3 by both RC and CD metrics. This can be explained by the unfortunate choice of datasets for carrying out these experiments. Selected dataset G_{orig} is extremely differ to other K_{train} and K_{prior}.

Therefore, the results of experiments show the possibility of adaptation external heterogeneous data sources for using in existing recommendation system. In this study, we succeeded in the unification of two extremely different data sources. Moreover, we managed the filtration and selection of the most useful pieces of G_{orig} data for adaptation to K_{train} dataset. However, we were not able to get a significant increase in the effectiveness of the original recommendation system. Thus, it is required to find more suitable datasets for such an experimental study. For the future work, we plan to: improve the process of data unification by using complex nature language processing models; build a clusterization model that can consider association rules as a distance between transaction vectors.

4. Conclusion

In this study, we explored the possibilities of improving the quality of a recommendation system for grocery supermarkets. We proposed the integration scheme that allows adapting external heterogeneous data sources for fit to existed recommendation system. The proposed integration scheme is based on several market basket analysis methods, such as association rules, collaborative-filtering and clusterization. For the experimental study, we found two datasets with transaction of data of supermarkets' customers. We define two metrics to evaluate the quality of constructed recommendation system and conduct experiments comparing the original and adapted recommendation systems. Our results show the ability to improve the quality of recommendation system by using additional heterogeneous data

sources. However, the selected datasets were extremely different and do not allow us to obtain the expected results. The main result of the study is that we understood the further ways of research that it is necessary to find more suitable datasets and improve the used methods and models.

Acknowledgment

This work financially supported by Ministry of Education and Science of the Russian Federation, Agreement #14.575.21.0165 (26/09/2017). Unique Identification RFMEFI57517X0165.

References

[1] M. Kaur and S. Kang, "Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining," *Procedia Comput. Sci.*, vol. 85, no. Cms, pp. 78–85, 2016.

[2] D. Solnet, Y. Boztug, and S. Dolnicar, "An untapped gold mine? Exploring the potential of market basket analysis to grow hotel revenue," *Int. J. Hosp. Manag.*, vol. 56, pp. 119–125, 2016.

[3] S. S. Umbarkar and S. Nandgaonkar, "Using Association Rule Mining: Stock Market Events Prediction from Financial News," *Int. J. Sci. Res. ISSN (Online Index Copernicus Value Impact Factor*, vol. 14, no. 6, pp. 2319–7064, 2013.

[4] A. Samecka-Cymerman, A. Stankiewicz, K. Kolon, A. J. Kempers, and R. S. E. W. Leuven, "Market Basket Analysis: A New Tool in Ecology to Describe Chemical Relations in the Environment-A Case Study of the Fern Athyrium distentifolium in the Tatra National Park in Poland," *J. Chem. Ecol.*, vol. 36, no. 9, pp. 1029–1034, 2010.

[5] D. I. Smith, M. F. Curran, and A. V. Latchininsky, "Market basket analysis of grasshopper (Orthoptera: Acrididae) assemblages in eastern Wyoming: a 17-year case study using associative analysis for ecological insights into grasshopper outbreaks," *Ecol. Entomol.*, vol. 42, no. 4, pp. 379–382, 2017.

[6] R. aa K. R. V and B. D. Jitkar, "Association Rule – Extracting Knowledge Using Market Basket Analysis," *Res. J. Recent Sci. ...*, vol. 1, no. 2, pp. 19–27, 2012.

[7] Z. Qureshi, J. Bansal, and S. Bansal, "A Survey on Association Rule Mining in Cloud Computing," vol. 3, no. 4, pp. 318–321, 2013.

[8] J. Nahar, A. B. M. S. Ali, T. Imam, K. Tickle, and P. Chen, "Brain Cancer Diagnosis-Association Rule-Based Computational Intelligence Approach," *2016 IEEE Int. Conf. Comput. Inf. Technol.*, pp. 89–95, 2016.

[9] P. Manchanda, A. Ansari, and S. Gupta, "The " Shopping Basket" A Model for Multicategory Purchase Incidence Decisions The " Shopping Basket" A Model for Multicategory Purchase Incidence Decisions THE " SHOPPING BASKET" A MODEL FOR MULTICATEGORY," *Source Mark. Sci.*, 1999.

[10] S. Kamley, S. Jaloree, and R. S. Thakur, "An Association Rule Mining Model for Finding the Interesting Patterns in Stock Market Dataset," *Int. J. Comput. Appl.*, vol. 93, no. 9, pp. 975–8887, 2014.

[11] P. B. Thorat, R. M. Goudar, and S. Barve, "Survey on Collaborative Filtering and Content-Based Recommending," *Int. J. Comput. Appl.*, vol. 110, no. 4, pp. 31–36, 2015.

[12] X. Yang, Y. Guo, Y. Liu, and H. Steck, "A survey of collaborative filtering based social recommender systems," *Comput. Commun.*, vol. 41, pp. 1–10, 2014.

[13] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet Comput.*, vol. 7, no. 1, pp. 76–80, 2003.

[14] Z. Lu, Z. Dou, J. Lian, X. Xie, and Q. Yang, "Content-based collaborative filtering for news topic recommendation," *AAAI 2015 Proc. Twenty-ninth AAAI Conf. Artif. Intell.*, pp. 217–223, 2015.

[15] M. Nilashi, O. Bin Ibrahim, and N. Ithnin, "Hybrid recommendation approaches for multi-criteria collaborative filtering," *Expert Syst. Appl.*, vol. 41, no. 8, pp. 3879–3900, 2014.

[16] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, 1993.

[17] D. Kajaree and R.. Behera, "A Survey on Machine Learning: Concept, Algorithms and Applications," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 5, no. 2, pp. 1302–1309, 2017.

[18] "Instacart Market Basket Analysis," *2017*. [Online]. Available: https://www.kaggle.com/c/instacart-market-basket-analysis.

[19] "Groceries dataset," *2014*. [Online]. Available: http://www.salemmarafi.com/wp-content/uploads/2014/.