Improvement of Retail Recommender System by Integration of Heterogeneous Sources of Data and Classification of Customers' Parameters

Mikhail Melnik, Tatiana Kutuzova
ITMO University
Saint-Petersburg, Russia
mihail.melnik.ifmo@gmail.com, kutuzova.tanya@mail.ru

***Abstract*—The growing assortment of goods in stores necessitates analysis of customer behavior to improve the quality of provided services. One of opportunities to improve the quality of service in the retail is to provide recommendations to customers. Recommender systems (RS) allow retailers to offer the most suitable sets of products for their customers that they might like to purchase. However, retailers do not always have enough information about customers' preferences to build a quality of RS. As a result, they need to expand their transaction databases through the use of external data sources. The Big Data Exchange can serve as a source of new data, which also provides opportunities for analysis and data expansion. Most often, data from various sources are heterogeneous, i.e. they are not presented in a single format and may contain different information about their clients or transactions. This leads to the need to transform data into a single form, and, consequently, to increase computational complexity of methods for data integration. Consequently, it is necessary to develop a heterogeneous data integration technique. Moreover, each client wants to get personalized recommendation which based not only on transaction history but also focused on their parameters such as age, marital status, income. However, not all data sources contain information about clients parameters. This study provides a classification of clients parameters for extending data by analyzing transaction history of initial data. This model allows user to achieve the better quality of RS, therefore, the higher profits for retailers and proper recommendations for clients.**

I. Introduction

With the increasing assortment in a retail, volumes of information which necessary to process also increases. In particular, with the rapid development of retail, there is a need to constantly improve the quality of provided services, for example, providing recommendations to the client.

Recommender systems (RS) allow retailers to offer their customers the most suitable sets of products that they might like to purchase. Thus, customers receive a recommendation, which may be a reminder to buy a certain set of products. To achieve the effect of recommendations, it is necessary to analyze the behavior of customers to build a high-qualitative RS. However, retailer may has not enough data to build such a system. To solve this problem, retailers can turn to the Big Data Exchange, where they can obtain external data from other sources of network retailers or a complete analysis of their own or expanded using other data sources. However, not all information may be useful to users of a Big Data Exchange (BDE). This is one of the main reasons why you need to extract important information from a large amount of data.

Big Data Exchange is a electronic platform that provides the use of corporate data provided by various providers and open source data. The project is currently being developed as a part of the program indicated in acknowledgment. This platform contains data mining methods and provides the following features for customers: expanding your database with external data; obtaining new and complete data segments; data analysis based on available in platform analytical tools. As a result of working with the BDE platform, an user can get completely new data from another area, expand their data from external sources or get data analysis. BDE can be used in various fields from retail to medicine or education.

Most often, data from various sources are heterogeneous, i.e. they are not presented in a single form, and various sources store different information about their customers. This creates a need to develop data analysis methods and integrate data from heterogeneous sources.

One of the goals of the project is to research and develop technologies to improve the quality of the retail recommendation system by integrating heterogeneous data from various sources within the framework of the Big Data Exchange. In particular regarding this work, we focus on the study of the possible characteristics of customers and their predictions for the development of integration technology. This can be useful in situations where one of the used datasets does not have customer information. In these cases, customer behaviour can be predicted by a model which based on the another dataset with known customers characteristics.

Furthermore, integration of heterogeneous data sources become a challenge and this study provides the following contributions:

- an analysis of demographic information to the most convenient data representation;
- an experimental comparison of different approach effectiveness for clients' parameters classification;
- an experimental research of ability recommender system improvement by integration of heterogeneous data.

ISSN 2305-7254

A. Related works

Actually, a retail recommender system (RS) in the context of market basket analysis (MBA) [1] provides an ability to get information about customers behavior. Recommender systems are widely used in different areas, such as finance, retail, and biology. Three main techniques for constructing RS existed: content-based, collaborative filtering and hybrid. A content-based technique is constructed on products properties and clients preferences. In the study [2] news topic recommender system is constructed by a content-based approach. Another example is the Netflix recommender system [17], which builds recommendations based on ratings from all users. Over the past few years, a number of unique e-shopping recommendation systems have been developed to provide guidance for individual customers on the Internet. Electronic shopping is a specialized and very popular area of electronic commerce [19]. Digital libraries are collections of digital objects, as well as related services provided to user communities [20]. Recommendation systems can be used in digital library applications to help users find and select sources of information and knowledge [21].

Collaborative filtering based on previous customers behavior. It comprises some techniques: model-based (recommendations based on the constructed model of customers behavior) and memory-based (recommendations constructed on the history of customers actions). Such methods can be found in Amazon [3] recommender systems. The construction of RS based on collaborative filtering in some situations can cause difficulties associated with the lack of sufficient data. For example, the problem of cold start in RS is solved in [17]. Moreover, the problem of insufficient data can be solved by integrating with external data of a similar nature. The hybrid approach combines the previous two, for instance, this approach is applicable for multi-criteria collaborative filtering in [4]. The integration of heterogeneous data sources process includes a unification of heterogeneous datasets, a clustering and a filtering of data for a recommender system construction. There is a wide range of different approaches but in this study, all data are transformed into a general form by a word embedding method. For instance, in the study [5] recommendation for a client based on the previous interaction with items constructed by word2vec. Moreover, [6] compares the effectiveness of item2vec and SVD approaches for RS. In the study [7] the library for analysis of clients behavior2vec was introduced and they constructed a recommendation based on the cosine distance between the distributed representations of the behaviors on items under different contexts. It is necessary to define only appropriate parts of existed data in datasets for integration to build a more effective and qualitative RS.

II. Methodology

The process of development of technology for improving the quality of the recommender system by analyzing characteristics of customers consists of the following stages:

- adaptation of heterogeneous data source and selection of external data;
- unification of selected data arrays;
- data integration;
- modelling of customer behavior;
- construction of improved recommender system based on integrated data;
- evaluation of a quality of built recommender system.

The scheme is presented in Fig. 1. Big Data Exchange is composed of system core, controller, data interface and data sources. Moreover, BDE includes all additional methods and analytical tools like a RS improvement technology as available services for BDE users.

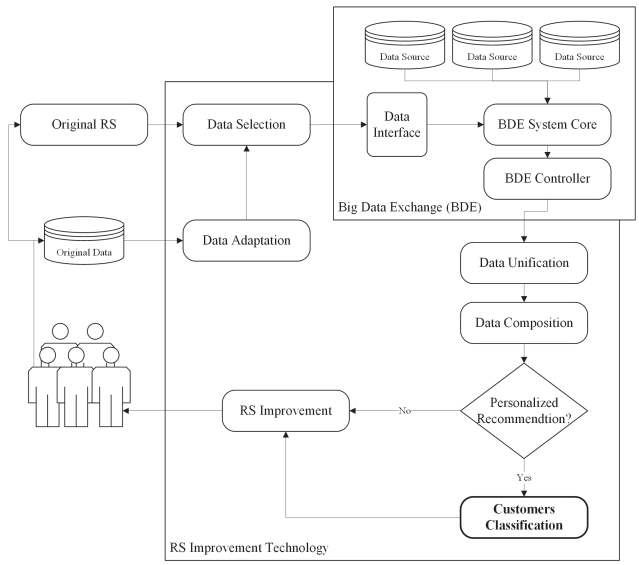

Fig. 1. Scheme of recommender system in framework of Big Data Exchange

A. Data unification

Data unification is an important part of integration methods in the context of BDE platform. It is necessary to transform heterogeneous retail datasets into a uniform view for the further joint analysis. Initial data processing is required because the data on the names of goods are presented in various forms according to some parameters: register, word endings, abbreviations, etc. Thus, there is a need to clear the data for further analysis. If changing the register is not a time-consuming task, then converting words from their current state to general or identifying abbreviations is a complex task.

After data cleaning and preprocessing [11] we calculate the distance between the vector representation of products names from internal and external data sources in order to determine the most appropriate pairs of products names from different sources. Three word embedding methods for transforming semantic data into a vector form in the context of data unification are compared in this paper:

- word occurrence(WO) [14] uses information about the frequency of words;
- latent semantic indexing(LSI) [15] transforms all products name into a vector form by using terms (set of unique words from all data sources);
- word2vec(W2V)[16] trains a neural network for a reconstruction semantic context of words.

TABLE I. COMPARING WORD EMBEDDING METHODS IN THE CONTEXT OF DATA UNIFICATION

	W2V	LSI	WO
Accuracy	64	25	3
Mean recall	0.85	0.82	0.64
Mean precision	0.61	0.59	0.31
Mean F1	0.71	0.69	0.42
Mean RC	0.54	0.49	0.32

To define the quality of unification there are marked test data by an expert. That marked dataset contains names with different length and complexity. The accuracy of matching by each method is presented in Table I. Clearly, unification by using W2V algorithm provide more effective results with accuracy 64% than other methods, especially WO method which accuracy is only 3%. Moreover, the quality of unification also affects the quality of RS. Mean values of metrics W2V and LSI are closed, although W2V allow achieving better quality RS. Low-quality unification method leads to low-quality RS.

B. Data integration

Data integration is the process of combining heterogeneous data sources to expand the database of network retail transactions in order to conduct a more accurate analysis of customer behavior. Data integration is based on the results of the unification process. The main component of the integration process is filtering unified external data.

Data integration is built on the vector representation of unified data. This allows combining different sources for improving the quality of RS. Data clustering is necessary to determine the most suitable data components for integration from external sources. Clustering can be based on various data, such as information about products, orders or customers, depending on the goal. Clustering of clients is carried out to select only the necessary groups of clients for the integration of transaction data.

Firstly, we use k-means for clustering of original data. This method is the most appropriate due to the ability to set a variable number of clusters and good performance with our data compared with other clustering methods, such as DBSCAN [18]. The value of silhouette coefficient 1 allows to measure the proximity of a sample to its own cluster. The value greater than zero shows that the sample has a strong connection with its own cluster and a weak connection with neighbouring clusters. The most proper number of clusters is calculated by finding the balance between mean and minimal values of silhouette coefficient.

$$s(i) = \frac{(b(i) - a(i))}{\max(a(i), b(i))} \tag{1}$$

where i sample for which silhouette coefficient is calculating, $a(i)$ average distance between i and other samples in the same cluster, $b(i)$ the minimum distance between i and samples from another clusters. Not all external orders are suitable for integration. Silhouette coefficient is also used for that. Initially, the closest original cluster is defined by calculating minimal Euclidian distance between external order vector and original orders vectors. After that, silhouette value is calculated for external order vector, if the value is more than zero the order should be integrated into original data, otherwise, it should be ignored.

C. Recommender system

We construct a RS based on two main collaborative filtering approaches: item-based and user-based. Slope One algorithm [8] is used as an item-based recommender system, where weight is the number of co-purchasing. For each product from customer basket N the most co-purchasing products are chosen. But this type of recommendation does not provide any personalization. Furthermore, we choose N the most popular products from transactions that were committed by customers with the same parameters. And the recommendation is a union of two algorithms recommendations without any products which customer has already has. As a result, this RS provides the recommendation that is convenient for current customer but do not only based on their statistics or similar customers but also based on common preferences of current market.

RS is an information filtering technology to predict the preference of users. Creating of RS consists of two main steps: association rules mining and similarity matrix calculation. For all products, matrix similarity is constructed. Each value is calculated with the following formula

$$sim(\vec{A}, \vec{B}) = \frac{\vec{A}\vec{B}}{\|\vec{A}\|\|\vec{B}\|} \tag{2}$$

where A, B vectors of products. This matrix represents the probability of pairs of products joint purchase. Moreover, this matrix allows expanding the list of initial products by adding the most co-purchasing products. Association rules (AR) mining is a part of the customers behaviour analysis. This process results in sets of products which are being co-purchased together by clients. The formal definition of mining AR has been described in [15]. AR includes two main metrics: support and confidence. Support is a part of transactions that contain all the elements of a set. The higher the support value, the more often a set of elements occurs. Associative rules with a high support value are most preferred for use in more transactions. Confidence is the probability that the AR contains certain elements in the both parts of a rule. The high value of this parameter demonstrates that the sets of products from a rule are acquired jointly than individually. RS provides determining of the most alluring recommendation for users. Recommendation compiled for a client with an initial list of products, which they are going to buy. Expanded by matrix similarity set of clients products are searched in the left part of AR, as a result, right part of determined rules is the most appropriate for the recommendation.

III. EXPERIMENTAL STUDY

There are three main experiments which oriented on RS improvement by customer classification in the context of heterogeneous data integration:

- data integration based on word embedding method and silhouette coefficient;
- classification of customers parameters;

- comparison of the quality of RS constructed on initial data and integrated data.

A. Data

We use two datasets from public access to conduct experiments with RS and heterogeneous data integration. First dataset is used as an initial data for training a classifier of client's parameters. Second dataset used as external for integration onto initial data.

Initial data. Dunnhumby data [9] were used as initial/original data. There is detailed information about clients and their transaction. There are 2595732 transactions, 276484 orders and 801 clients. Orders contain information about 98856 products. Clients' description contains 7 groups which contain some subgroups:

1) AGE - the age range of customers [6 subgroups from 19 to 65+]
2) MARITAL STATUS - the marital status customers [3 subgroups: married, not married, unknown]
3) INCOME - the annual income of customers [12 subgroups from Under15K to 249K]
4) HOMEOWNER - groups of home ownership [5 subgroups: homeowner, probable owner, renter, probable renter, unknown]
5) HH COMP - the category of people/family [6 subgroups: 2 adults no kids, 2 adults kids, 1 adult kids, single female, single male, unknown]
6) HOUSEHOLD SIZE - the amount of people who live in the house [5 subgroups from 1 to 5+]
7) KID CATEGORY - the amount of kids [4 subgroups: 1, 2, 3+, unknown]

Transactions of initial data have the form as in Table II.

External data. External data were got from "Instacart Market Basket Analysis" [10] kaggle competition. There is no information about customers except of their identification number although there are 1384617 transactions, 131209 orders and 131209 unique customers. Transactions of external data have another form; the example is presented at Table III.

From tables II and III clearly that data sources are heterogeneous due to different format of data, especially product names.

B. Evaluation metrics

We have a prior model which consists of more complete information about customers behaviour and used for evaluation of another RS. Evaluation of constructed RS represents a comparison of prior and constructed recommendations. There are various metrics [16] of recommender systems evaluation but most of them are not suitable without any clients ratings.

Let we have a recommender system which can be described by the following equation:

$$R(c, b) = r \tag{3}$$

where c is client's parameters, b a current buying order and r is the personalized recommendation to a client with c parameters. Moreover, there are a validation model of recommender system P which can be described as:

$$P(c, b) = p \tag{4}$$

where p is a validation recommendation to a client with c parameters. Each recommendation is a set of products which are the most suitable for a customer. A great amount of different metrics for RS are existed but most of them based on client's grading. In retail RS there are no customers' grades as a results it is impossible to use that metrics. Although, there are some metrics which allow to evaluate a set of recommendations: $F1$ 5 and Jaccard index J 6.

$$F1(R) = \frac{2(p \cap r)}{p \cup r} \tag{5}$$

$$J(R) = \frac{r \cap p}{r \cup p} \tag{6}$$

C. Results

We provide a three sets of experiments to achieve a RS improvement:

- integration of heterogeneous data sources;
- classification of clients' parameters;
- recommender system construction.

a) Integration of heterogeneous data sources: Two data sets contain information in different form, as a result it is impossible to use them without any transformation. We use fasttext [11] method for word embedding. Data integration technique allows to transform all data to unified form and use it together. Initial data were divided by k-means [12] on 90 clusters due to the higher value of average silhouette score [13] for initial data. The distribution of silhouette coefficient for external data is presented in Fig. 2. A lot of data hold on the border with values close to zero, although all data which silhouette coefficient is more than zero should be integrated. After all, steps of data integration about 42% of external products will use in further experiments.

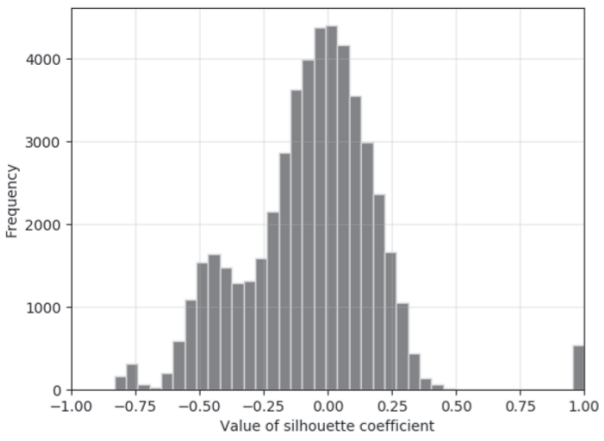

Fig. 2. The distribution of external data silhouette coefficient

After transaction data integration there are a problem that not all our data contain information about clients. But we

TABLE II. EXAMPLE OF INITIAL DATA TRANSACTIONS FROM DUNNHUMBY DATASET

household_key	BASKET_ID	PRODUCT_ID	COMMODITY_DESC	SUB_COMMODITY_DESC
1	26984851472	1004906	POTATOES	POTATOES RUSSET (BULK&BAG)
7	27165356331	1033142	ONIONS	ONIONS SWEET (BULK&BAG)
8	27190457873	1036325	VEGETABLES - ALL OTHERS	CELERY

TABLE III. EXAMPLE OF EXTERNAL DATA TRANSACTIONS FROM KAGGLE DATASET

order_id	user_id	product_id	product_name
1	112108	47209	Organic Hass Avocado
36	79431	46620	Cage Free Extra Large Grade AA Eggs
36	79431	39612	Grated Pecorino Romano Cheese

integrate only useful information which is so close with initial data thus clients should be similar too.

b) Classification of clients' parameters: The target set of experiments was carried out, aimed at predicting user parameters in order to achieve an improvement in the quality of the recommendation system. Data contain seven clients' parameters which describe age, marital status, income and etc. We test four approaches to classify N the values of these parameters. Vectorized transaction data was used to classify customer parameters. Input data are the vectors of goods that were purchased by customers. The output is the parameter vector of the customers who purchased the input data item. For classification we used Decision Tree algorithm.

The first approach (Exp 2.1) provides a using of all data about clients which we have. The second approach (Exp 2.2) suggest to remove clients with 'unknown' parameters because that clients can be belong to each class. Exp 2.3 provides that the most frequently confused classes should be merge. And the last is Exp 2.4 merges the most similar classed based on transactional statistics. For instance, if people from age classes 19-24 and 25-30 buy the same products than these classes can be merged.

From Table IV it is clear that the last approach with merging the most similar groups of clients to reduce the number of classes should be use to classify parameters AGE, INCOME, HOMEOWNER and HOUSEHOLD SIZE and the removing clients with unknown parameters more suitable for MARITAL STATUS and HH COMP, for KID CATEGORY it is not necessary to reduce any classes but the reducing the number of classed based of classifiers error does not convenient for any parameters.

c) Recommender system construction: Results from previous two experiments are used for RS improvement. Two recommender system were constructed based on algorithm which described in Section 2. The first based on initial data (D_{init}) and the second based on extended integrated data (D_{inter}). Clients classification based on integration of heterogeneous data allows to improve RS by 44% based on $F1$ metric and by 45% on Jaccard index in compare to original D_{init} RS. Distribution of values F1 and Jaccard metritis are presented in Fig. 3.

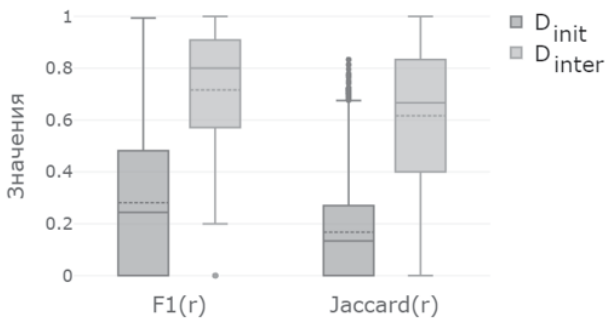

Fig. 3. Comparison of qualities of two recommender system with and without analysis of clients' parameters

IV. CONCLUSION

Recommender system is a vital service for retail field. This is due to the better recommendation they made to customers the more profits they can get. This paper presents a technology for improvement of recommender system in the retail field based on integration of heterogeneous data sources within the Big Data Exchange. In the course of the work, the approaches to analyzing the market basket of network retail for providing services within the Big Data Exchange were investigated, as well as methods for bringing heterogeneous retail data into a single form (data unification). Further, methods for unifying several heterogeneous data sources using word embedding methods were implemented, as well as clustering of external data to extract the most useful data for integration. All this methods were composed as a technology for integrating of heterogeneous data sources. As a result, technology for integrating heterogeneous data from various sources can improve the quality of the recommender system by expanding data on customer behavior; within the framework of the experiments, the improvement of the recommendation system was at least 20%. As part of the integration of heterogeneous retail data, it is possible to classify user parameters if they are not available in the external data or are presented in a different form to improve the quality of the constructed recommendation system. For that we investigated classification of clients from external data based on their purchasing. The extended recommender system based on customer classification achieved about 45% quality improvement in compare to original one.

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.

TABLE IV. COMPARISON THE ACCURACY OF CLIENT CLASSIFICATION APPROACHES

	AGE	MARITAL STATUS	INCOME	HOMEOWNER	HH COMP	HOUSEHOLD SIZE	KID CATEGORY
Exp 2.1	0.42	0.50	0.28	0.67	0.37	0.42	0.68
Exp 2.2	0.45	0.76	0.27	0.90	0.85	0.46	0.47
Exp 2.3	0.68	0.49	0.28	0.67	0.39	0.42	0.68
Exp 2.4	0.80	0.25	0.79	0.96	0.56	0.86	0.63

REFERENCES

[1] *H*eydari, Majeed, and Amir Yousefli. "A new optimization model for market basket analysis with allocation considerations: A genetic algorithm solution approach." *Management and Marketing* 12.1 (2017): 1-11.

[2] *L*u, Zhongqi, et al. "Content-based collaborative filtering for news topic recommendation." *Twenty-ninth AAAI conference on artificial intelligence*. 2015.

[3] *L*inden, Greg, Brent Smith, and Jeremy York. "Amazon. com recommendations: Item-to-item collaborative filtering." *IEEE Internet computing* 1 (2003): 76-80.

[4] *N*ilashi, Mehrbakhsh, Othman bin Ibrahim, and Norafida Ithnin. "Hybrid recommendation approaches for multi-criteria collaborative filtering." *Expert Systems with Applications* 41.8 (2014): 3879-3900.

[5] *O*zsoy, Makbule Gulcin. "From word embeddings to item recommendation." *arXiv preprint arXiv* 1601.01356 (2016).

[6] *B*arkan, Oren, and Noam Koenigstein. "Item2vec: neural item embedding for collaborative filtering." 2016 *IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2016.

[7] *C*hen, Hung-Hsuan. "Behavior2Vec: Generating Distributed Representations of Users Behaviors on Products for Recommender Systems." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 12.4 (2018): 43.

[8] *B*asu, Anirban, Jaideep Vaidya, and Hiroaki Kikuchi. "Efficient Privacy-Preserving Collaborative Filtering Based on the Weighted Slope One Predictor." *J. Internet Serv. Inf. Secur.* 1.4 (2011): 26-46.

[9] *D*unnhumby data [Electronic resource]. URL: https://www.dunnhumby.com/careers/engineering/sourcefiles.

[10] *I*nstacart Market Basket Analysis [Electronic resource] // 2017. URL: https://www.kaggle.com/c/instacart-market-basket-analysis.

[11] *P*orter, Martin F. "Snowball: A language for stemming algorithms." (2001).

[12] *H*artigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979): 100-108.

[13] *R*ousseeuw, Peter J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics* 20 (1987): 53-65.

[14] *T*atiana, Kutuzova, and Melnik Mikhail. "Market basket analysis of heterogeneous data sources for recommendation system improvement." *Procedia Computer Science* 136 (2018): 246-254.

[15] *H*ofmann, Thomas. "Probabilistic latent semantic indexing." *ACM SIGIR Forum.* Vol. 51. No. 2. ACM, 2017.

[16] *C*hurch, Kenneth Ward. "Word2Vec." *Natural Language Engineering* 23.1 (2017): 155-162.

[17] *W*ei, Jian, et al. "Collaborative filtering and deep learning based recommendation system for cold start items." *Expert Systems with Applications* 69 (2017): 29-39.

[18] *K*umar, K. Mahesh, and A. Rama Mohan Reddy. "A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method." *Pattern Recognition* 58 (2016): 39-48.

[19] *L*i, Seth Siyuan, and Elena Karahanna. "Online recommendation systems in a B2C E-commerce context: a review and future directions." *Journal of the Association for Information Systems* 16.2 (2015): 72.

[20] *G*onalves, Marcos Andr, et al. "Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries." *ACM transactions on information systems (TOIS)* 22.2 (2004): 270-312.

[21] *P*orcel, Carlos, and Enrique Herrera-Viedma. "Dealing with incomplete information in a fuzzy linguistic recommender system to disseminate information in university digital libraries." *Knowledge-Based Systems* 23.1 (2010): 32-39.