COMER: ClOud-based MEdicine Recommendation

Yin Zhang†, Long Wang†, Long Hu†, Xiaofei Wang♮, and Min Chen†
† Departent of Computer Science and Technology, Huazhong University of Science and Technology
1037 Luoyu Road, Wuhan, 430074, China
♮ Department of Electrical and Computer Engineering, The University of British Columbia
Vancouver, BC, Canada V6T 1Z4
Email: yin.zhang.cn@ieee.org, longwang.epic@gmail.com, longhu.cs@gmail.com, xfwang@ece.ubc.ca,
minchen2012@hust.edu.cn

***Abstract*—With the development of e-commerce, a growing number of people prefer to purchase medicine online for the sake of convenience. However, it is a serious issue to purchase medicine blindly without necessary medication guidance. In this paper, we propose a novel cloud-based medicine recommendation, which can recommend users with top-N related medicines according to symptoms. Firstly, we cluster the drugs into several groups according to the functional description information, and design a basic personalized medicine recommendation based on user collaborative filtering. Then, considering the shortcomings of collaborative filtering algorithm, such as computing expensive, cold start, and data sparsity, we propose a cloud-based approach for enriching end-user Quality of Experience (QoE) of medicine recommendation, by modeling and representing the relationship of the user, symptom and medicine via tensor decomposition. Finally, the proposed approach is evaluated with experimental study based on a real dataset crawled from Internet.**

***Index Terms*—Cloud; QoE; Medicine Recommendation; Collaborative Filtering; Clustering; Tensor Decomposition**

I. Introduction

With the development of information technology, the most advanced technologies are implemented in the field of health care [1] [2]. With the popularization of medical knowledge and the explosion of e-commerce, a growing number of people try to purchase medicine via online pharmacies, such as Drugstore [3]. Because of no limitation in terms of time, space and region, online pharmacies are more convenient than the traditional pharmacies. In addition, customers can get detailed information of drug and user rating on line. Best of all, because online pharmacies can reduce the cost of storage, sales and employees, the price of medicine will be cheaper than the normal pharmacies generally. However, there are following issues of purchasing medicine online in spite of its convenience:

- *Validity*: There is a high risk of buying the noneffective even wrong medicine because of the illegal and exaggerated information on the website.
- *Reliability*: The most challenge is purchasing medicine online without professional instruction. The nonprofessionalities, such as ignoring the interaction of multiple medicines, using medicine repeatedly, etc., would cause serious consequence for the patients.

In order to provide personalized healthcare services, several recommendation technologies are implemented in heath care applications. In [4], Duan et al. use correlations among nursing diagnoses, outcomes and interventions to create a recommender system for constructing nursing care plans. In [5], Kim et al. develop a personalized u-healthcare service system providing item recommendation based on context-aware model. In [6], Chen et al. design a diet recommendation system which has the expert knowledge of three high chronic diseases and recommend suitable foods for users according to their health information.

Although these earlier works demonstrate the value of personalized healthcare service, most of them recommend healthcare just from low-dimensional data, especially pharmaceutical and commercial data. In this paper, we introduce some algorithms of personalized recommendation system and analyze the advantages and disadvantages of recommendation based on collaborative filtering algorithm. Furthermore, we propose a novel cloud-based medicine recommendation (COMER) to help patients to find accurate medicine on the Internet. More specifically, this paper makes the following contributions in COMER.

- We propose a basic COMER based on clustering and collaborative recommendation algorithms to provide customers with sufficient guidelines.
- We improve the basic COMER with implementation of tensor decomposition to support validate and professional COMER based on massive and sparse drug data.

The remainder of this paper is organized as follows. We present significant proposed personalized recommendation technologies and healthcare integration network proposal in Section II. In Section III, we introduce K-means clustering and tensor decomposition, which are the key mathematical methods used in COMER. Section IV describes our proposal for medicine recommendation. Section V shows the performance of our proposal via experimental results and analysis comparing with collective filtering algorithm. Finally,

978-1-63190-025-9 © 2014 ICST
DOI 10.4108/icst.qshine.2014.256542

Section VI concludes this paper.

II. Related Work

With the increase of information on the internet, the problem of information overload is becoming more and more serious, thus the personalized recommendation system [7] [8] [9] are emerged to help people find the content of interest much more fast.

The e-commerce recommendation system is defined as: *use the e-commerce sites to provide product information and recommendations to clients, thus to help users decide what goods to buy, which simulates the process of sales assistance of purchasing goods to customers*.

The personalized recommendation technology is the core technology in electronic commerce recommendation system, which determines the effect and the performance of the recommendation system.

The collaborative filtering recommendation is one of the most studied personalized recommendation technology, which uses the similarity based on users or items to recommend. In [10], Barragns-Martłnez et al. use collaborative filtering recommendation mixed with singular value decomposition algorithm to recommend TV program and has reached an unprecedented effect. In [11], Cai et al. use collaborative filtering algorithm to do the personalized recommendation for social network users, and make accurate predictions for a person's dating tendency.

The content-based recommendation is to learn the user's interest based on the characteristics of the user's past behaviors, and do the recommendation according to the matching degree of the user's interest and the items to be predicted, such as the news recommendation system. In [12], Lops et al. have done some overview introduction of recommendation system based on items' content.

The knowledge based recommendation usually need the support of knowledge in the certain field related to the items. In [13], Carrer-Neto et al. use the expert knowledge in the field of films to do some personalized recommendation of the movies that users may be interested in and have achieved a good result.

The association rule-based recommendation usually appears in transaction data analysis, which uses the association rules to discover the inner link of recommended objects and do the personalized recommendation. The discovery of association rules is very time-consuming, but can be solved by off-line data analysis. In [4], Duan et al. analyze the patient's information by using the correlation data mining algorithms, and recommend the best treatment options for patients. This is a new try of recommendation algorithm in the medical field, which is very promising.

III. Preliminaries

A. K-means clustering algorithm

Clustering is a method of data mining [14], which is commonly used in data analysis; K-means [15] is also a clustering method that is widely used, the goal is to divide the data points into k clusters, find the center of each cluster, and minimize the function

$$\arg_s \min \sum_{i=1}^{k} \sum_{x_j \in s_i} ||x_j - u_i||^2 \tag{1}$$

where u_i is the center of the cluster. This function requires each data point to be as closer as possible to the cluster center they belong to.

In order to get the center of each cluster, K-means performs two operations iteratively. Firstly, it gives some k centers randomly, and then classifies each data point to its nearest center, so we will construct the k cluster.

However, it is obvious that the k center position is not very accurate because of randomness. So it needs to bring the cluster center to the average position of the internal data points which are already got. In fact, the main task is to calculate the extreme value of the above function under the circumstances that each data point is classified already, and then construct a new k clusters.

In this process, the position of the center point changes constantly, and the structure of cluster is changing too. Through many times of iteration, the k centers will eventually stop moving and converge.

B. Tensor and tensor decomposition

Tensor is a spatial representation of a multidimensional array, which behaves as the N-dimensional vector space. For example, 3rd-order tensor can represent three kinds of data, and each coordinate can represent one kind of data [16]. Matrix is represented by two-dimensional data, which can be also comprehended as two-dimensional tensor. When there is more than three-dimensional tensor data, it will be called the higher-order tensor.

Through the tensor decomposition, we can use the values in the three-dimensional tensor in different applications, such as the field of predictive and personalized recommendation.

Tucker tensor decomposition: Tucker tensor decomposition is a principal component analysis method of the high dimension data, and it decomposes the original tensor into the product of a core tensor and a series of matrixes. Tensor decomposition can do the dimension reduction in the process of computing the core tensor.

Tucker decomposition theorem: a tensor A $\in R^{I_1 \times \cdots \times I_N}$ can be expressed as

$$A = B \times U^{(1)} \times U^{(2)} \times \cdots \times U^{(N)} \tag{2}$$

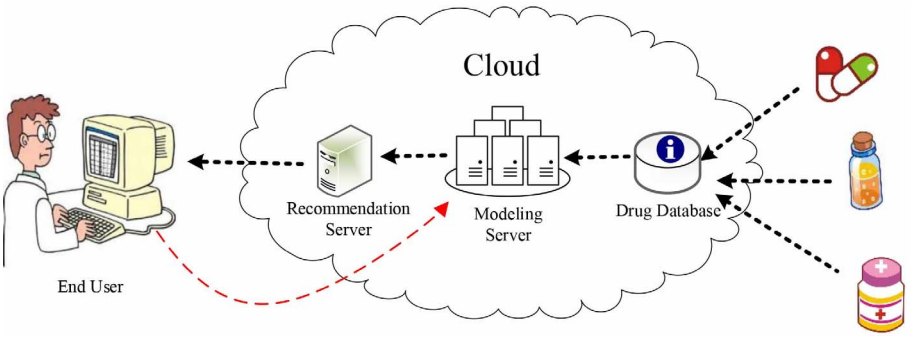

Fig. 1. Architecture of COMER

where B is the core tensor; $U^{(1)}, U^{(2)}, \cdots, U^{(N)}$ are a series of orthogonal matrixes. Due to the orthogonality of the projection matrix U, core tensor B can be obtained by the above equation

$$B = A \times U^{(1)T} \times U^{(2)T} \times \cdots \times U^{(N)T} \tag{3}$$

One of the most successful application fields of tensor model is the personalized label recommendation system, in which the original collaborative filtering approach is lack of scalability when faced the three-dimensional or more and cannot reflect the complex relationship among multi-dimensional data effectively, but the tensor model is a good solution to these problems.

IV. Design of COMER

In Figure 1, it presents the architecture of COMER. In Cloud, there are three important components.

- *Drug Database*: The detail information of various drugs have been collected from online drug store and social networks. In the database, it doesn't only store all the drug data, but also preprocesses the drug data, such as cleaning data, clustering drugs according to their feature, indications and other characteristics.
- *Modeling Server*: In modeling server, according to drug details and customer rating, we establish a offline recommendation model based on tensor. While a new drug or customer rating updates, the models will be rebuilt.
- *Recommendation Server*: While the end user inputs some keywords, such as drug name, symptoms, etc., recommendation server finds top-N relative drugs in drug database through recommendation model.

In the progress of medicine recommendation, we firstly use the vector space model (VSM) [17] to format the drug character according to the description of drug information and K-means algorithm to cluster drugs, and then use the user evaluation to do the collaborative filtering recommendation. After that, according to the shortage of the collaborative filtering recommendation, we propose the medicine recommendation model based on tensor decomposition.

A. Drug Clustering

Firstly we cluster drugs according to its treatment efficacy, the purpose of clustering is classify these drugs into different groups, then we can provide the essential drugs list that corresponds with the condition that the user input, which will facilitate personalized recommendation in the next step.

In this paper, we use the vector space model that is commonly used in information retrieval to represent each drug. Its main idea is: each of the drugs are mapped into a point of the vector space by a standardized set of orthogonal vectors. All drugs can be represented by terms vector (T_1,W_1,T_2,W_2,...T_n,W_n)in the space(T_i is feature word; W_i is the weight of T_i). Here we consider the feature vector as the subset of disease symptoms that this drug can cure.

First of all, we should count all the terms of the medicine description, and take them as a dictionary after removing all the stop words, then calculate the weight of drug words which appeared in the dictionary by using the method tf-idf [18] to obtain the representation of each drugs feature vector. The tf-idf which corresponds to the jth word in the dictionary and in the description of the j^{th} drug is:

$$TF{-}IDF(t_k, d_j) = TF(t_k, d_j) \cdot \log \frac{N}{n_k} \tag{4}$$

where TF(t_k,d_j) is the number of times of the k^{th} word in the description of drug j, and n_k is the drug number of all the drug descriptions include the k^{th} word.

After obtaining the feature vector of each drug, we cluster the drugs by using the k-means algorithm, and set the sample clustering into k clusters. The pseudo code of the drug clustering is given in Algorithm 1.

B. Drug Collaborative Recommendation based on User Scoring

The clustering in the above step is to prepare data for collaborative recommendation in the following. Every time when an user input the illness symptom, he can get some closest drug classifications according to the clustering and obtain a drug list which is most suitable to the users needs. In the base of the drug list, we extract all the users scoring of the

Algorithm 1 Drug Clustering

begin
Randomly select k cluster centroid points as $u_1, u_2, \cdots, u_k \in R^n$
loop
for each sample i
loop
for each cluster j, calculate the distance between sample i and cluster j: d_j=distance($x^{(i)}$,u_j)
end loop
select the cluster that sample i belongs to based on the $d_j, j = 1, \cdots, k$
loop
for each cluster j, recalculate the centroid of the cluster
end loop
the iteration count n plus one
end loop

drugs and construct a User - Drug scoring matrix , and then do the collaborative recommendation by the scoring matrix as following:

1) *User Similarity Calculation*: Calculate similarity between users(we use Pearson distance) is shown in Equation (5):

$$sim(a,b) = \frac{\sum_{i \in I}(R_{a,i} - \overline{R_a})(R_{b,i} - \overline{R_b})}{\sqrt{\sum_{i \in I}(R_{a,i} - \overline{R_a})^2 * \sum_{i \in I}(R_{a,i} - \overline{R_b})^2}} \tag{5}$$

where I is the set of all the items, $R_{a,i}$ and $R_{b,i}$ represent the score of the object i which are given by user a and user b respectively, $\overline{R_a}$ and $\overline{R_b}$ represent the average score of user a and user b respectively.

2) *Calculation of Predicted Value*: We predict the items that havent been scored by the given user according to the similarity between users that has been calculated before, and here we use the method of weight sum. We do the weight sum to the score of object i which is given by the similar users of the given user u, while the weight is the similarity of user u and each similar user, and calculate the score of object i of the given user u. The equation is shown in Equation (6).

$$P_{u,i} = \frac{\sum_{j}^{N}(s_{u,j} * R_{j,i})}{\sum_{j}^{N}(|s_{u,j}|)} \tag{6}$$

where N is the number of similar users, $S_{u,j}$ is the similarity of user u and user j, $R_{j,i}$ is the score of object i given by user j. Calculate all the drugs that havent been scored by the given user and recommend the top-N drugs that have the highest predicted score.

However, with the rapid growth of user number, the recommendation system input data set also increased and there comes a big challenge to the collaborative filtering recommendation algorithm. Also the collaborative filtering cannot solve the cold start problem and data sparsity problem, so we proposed the personalized medicine recommendation model based on tensor decomposition.

C. Recommendation based on Tensor Decomposition

In order to overcome the problem above, we propose the medicine recommendation algorithm based on tensor decomposition, which use the "User-Item-Tag" three tuple to model an third-order tensor. And then users receive personalized recommendations based on extracted core tensor according to the drugs predicted score. The main algorithm flow is shown below, wherein the data pre-process is completed in the previous stage.

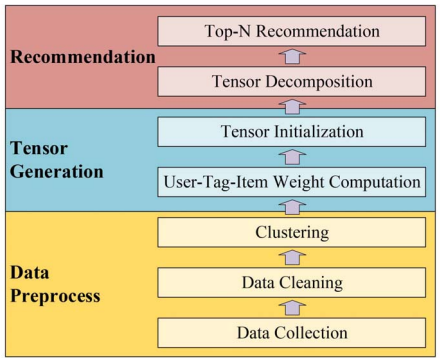

Fig. 2. Model flow chart

1) Tensor Modeling: In this paper, we propose a weight tensor representation based on users scoring, considering the weights of user, drug, drug tag and user scoring. We obtain a weight $w_{u,i,t}$ of "User - Item - Tag" three tuple, then when constructing the tensor model, we can take the weight $w_{u,i,t}$ as one of the elements of tensor to do the modeling.

The weight $w_{u,i,t}$ is calculated through the following method, first of all, cluster all the drugs by the K-means algorithm that mentioned in the above chapter, each of the drugs is mapped into a point of the vector space by a standardized set of orthogonal vectors, to the given users, drugs, drug tags, using the weight of labels in each drug to multiply the score of drug given by the user, we can obtain the $w_{u,i,t}$.

We define U as user set, I as drug set, T as label set of a certain drug, $R_{i,j}$ as the score of drug j given by user i, $T_{i,j,k}$ as the weight of user i , drug j , tag k in the feature vector, and $\sum_k T_{i,j,k} = 1$. $w_{i,j,k}$ represents the weight of the given user, drug and the drugs certain label in three-order tensor, $w_{i,j,k} = R_{i,j} \bullet T_{i,j,k}$, $i \in U, j \in I, k \in T$. So we have the tensor model as shown in Figure 3.

2) Tensor Decomposition and Recommendation: Here we get some new elements by the tensor decomposition according to the tensor decomposition method. The new elements represent the predict score that the users give to the drugs according to certain tags, so we can use the two dimensions of "User-

Tag" in the three-order tensor to recommend the top-N drugs which have the highest scores.

As shown in Figure 4, some new elements change from zero to the nonzero through the tensor decomposition, according to the weight of new elements obtained in the picture, we can provide the Top-N medicine recommendation list to the given users.

V. Experimental Results and Analysis

A. Data Source

The dataset has been crawled from the online drug store walgreens (http://www.walgreens.com/) in March, 2014, including 21,559 evaluation records of 8,163 drugs from 6,071 customers. (This dataset can be dowload at: http://www.datatang.com/data/46261) The attributes of each data record includes drug name, drug description, user name, gender, age, and user score. Data has been preprocessed at first, and the datasets after removing the shortened data is shown in Table I:

B. Experimental Evaluation Standard

This paper uses the accuracy and recall rate as the evaluation index, which are most commonly used in the recommendation system. In the progress of experiment, we take some users' "user-drug-tag" three tuple as the test sets, for each user, we can get the Top-N medicine recommendation list which is obtained by tensor decomposition, then we check whether the drug is in the recommendation list with the given users and tags, accuracy calculation function is presented in Equation (7):

$$Precision = \frac{\sum_{i=1}^{M} \frac{hitmedicine_i}{N}}{M} \tag{7}$$

where M is the predicted user number, $hitmedicine_i$ is the three tuple number(drugs corresponded with corresponding labels) of the recommendation list.

Recall rate calculation function is shown in Equation (8)

$$Recall = \frac{\sum_{i=1}^{M} \frac{hitmedicine_i}{H_i}}{M} \tag{8}$$

where M is the predicted user number,$hitmedicine_i$ is the triple number(drugs corresponded with corresponding labels) in the recommendation list, H_i is real triple number of each user.

Finally, we use F1 evaluation method to weigh the evaluation recommendation results effect wholly as shown in Equation (9).

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{9}$$

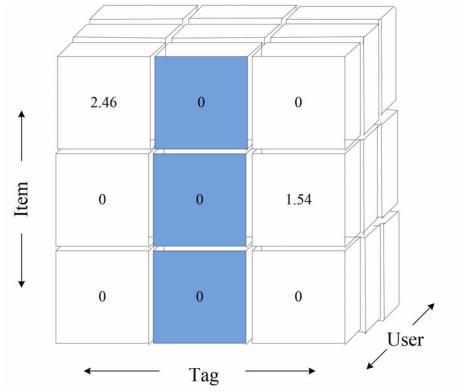

Fig. 3. "User-Item-Tag" three tuple tensor

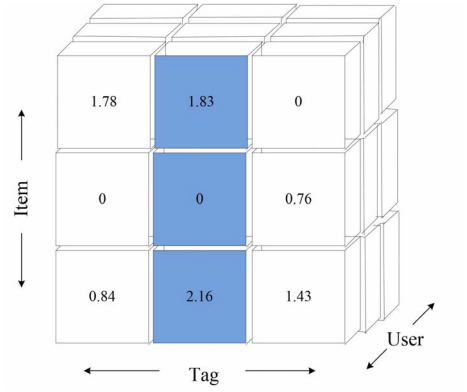

Fig. 4. After tensor decomposition

C. Results and Analysis

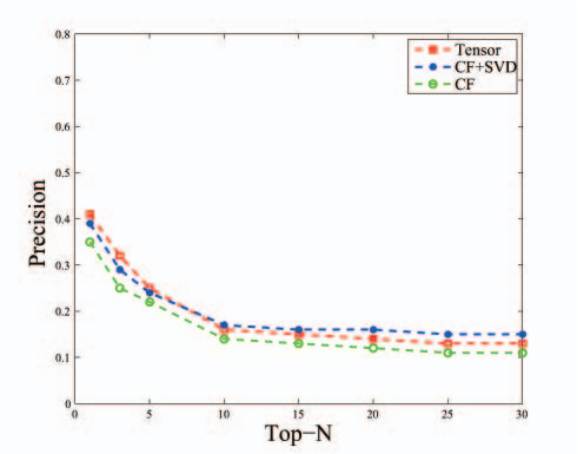

Fig. 5. The accurate rate of Tensor and CF

In this paper, we mainly use Python language and Matlab platform to implement the collaborative filtering and tensor decomposition algorithm.

TABLE I
Model Energy Consumption Parameter

Data Set	Drug Evaluation Number	User Number	Drug Number
Original Data	21,559	6,071	8,163
After pretreatment	20,397	5,827	7,901

Firstly, we cluster the drugs according to drug feature vector got from drug description, and here we take the cluster number as 20. The elements in the feature vector are expressed as symptom tags in accordance with each user and each drug. We selected 2000 evaluation data randomly as the test set in advance, while the rest of the data are as the training set.

In the progress of users' collaborative filtering recommendation, we first select all the relevant user - drug evaluation data according to test set users symptoms tags and construct a score matrix of "user-drug" who have evaluated these drugs, Considering the sparsity of the matrix, here we use the SVD method to reduce the dimension of the matrix, in the experiment part, we compared the result of whether use SVD or not. After we get the evaluate score matrix, we can calculate the similarity of the other users with the given user, and then use 10 most similarity users to calculate all the drugs which havent been scored by the given user, then recommend the top-N drugs with the highest predicted score.

We choose the core tensor of $core = (6*6*6)$ to decompose in the tensor decomposition, then we keep the element more than 0.1 in the approximate tensor we have got.

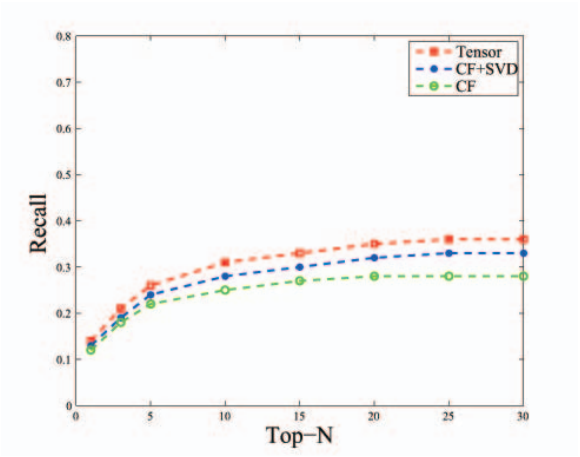

Fig. 6. The recall rate of Tensor and CF

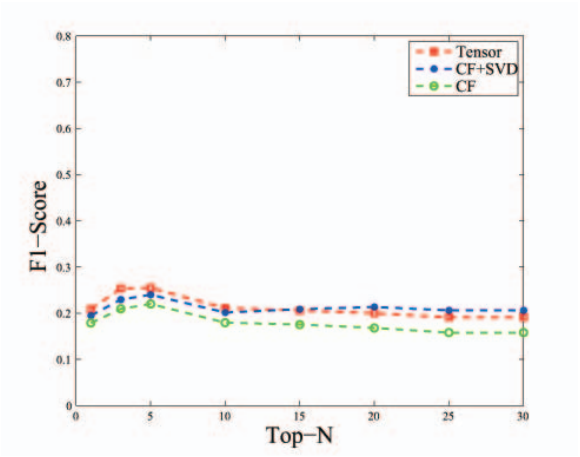

Fig. 7. The F1 index of Tensor and CF

According to the different length of recommendation list 1, 3, 5, 10, 15, 20, 25, and 30, we get the comparison of experimental results of collaborative filtering (CF), collaborative filtering(CF) with SVD and tensor decomposition (Tensor).

Experimental results show that, the tensor recommendation accuracy is higher than collaborative filtering when the recommend length is less than 10. And the accuracy is decreasing as the denominator increases obviously.

From the experimental results, we can see that when dealing with large amounts of sparse data, the recommendation list obtained based on tensor decomposition is better than collaborative filtering algorithm; but when the recommendation is more than 10, the accuracy declines quickly due to the proximity of the new elements after the tensor decomposition.

VI. Conclusions

In this article, we have proposed an approach with clustering and collaborative recommendation algorithms to meet the challenge of purchasing medicine online without sufficient instruction. On this basis, we have proposed COMER based on tensor decomposition considering the shortage of collaborative filtering when dealing with massive and sparse medicine data. Furthermore, we have shown that COMER can provide a valid, reliable, and effective medicine recommendation according to each customer's demand.

For future work, we will investigate how to improve the accuracy of COMER via combination with user's more characters such as age, geography and other factors.

References

[1] G. Hripcsak and D. J. Albers, "Next-generation phenotyping of electronic health records," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 117–121, 2013.

[2] R. Deshpande, W. Thuptimdang, J. DeMarco, and B. J. Liu, "A collaborative framework for contributing dicom rt phi (protected health information) to augment data mining in clinical decision support," in *SPIE Medical Imaging*. International Society for Optics and Photonics, pp. 90 390K–90 390K, 2014.

[3] Drugstore. Available: http://www.drugstore.com.

[4] L. Duan, W. N. Street, and E. Xu, "Healthcare information systems: data mining methods in the creation of a clinical recommender system," *Enterprise Information Systems*, vol. 5, no. 2, pp. 169–181, 2011.

[5] J. Kim, K. Y. Chung, "Ontology-based healthcare context information model to implement ubiquitous environment," *Multimedia Tools and Applications*, pp. 1–16, 2013.

[6] R.-C. Chen, Y.-D. Lin, C.-M. Tsai, H. Jiang, "Constructing a Diet Recommendation System Based on Fuzzy Rules and Knapsack Method," *Recent Trends in Applied Artificial Intelligence*, pp. 490–500, 2013.

[7] A. Jøsang, G. Guo, M. S. Pini, F. Santini, and Y. Xu, "Combining recommender and reputation systems to produce better online advice," in Modeling Decisions for Artificial Intelligence. Springer, pp. 126–138, 2013.

[8] Y. S. Cho, S. C. Moon, S.-p. Jeong, I.-B. Oh, and K. H. Ryu, "Clustering method using item preference based on rfm for recommendation system in u-commerce," in Ubiquitous Information Technologies and Applications. Springer, pp. 353–362, 2013.

[9] X. Yuan, J.-H. Lee, S.-J. Kim, and Y.-H. Kim, "Toward a user-oriented recommendation system for real estate websites," *Information Systems*, vol. 38, no. 2, pp. 231–243, 2013.

[10] A. B. Barragáns-Martínez, E. Costa-Montenegro, J. C. Burguillo, M. Rey-López, F. A. Mikic-Fonte, and A. Peleteiro, "A hybrid content-based and item-based collaborative filtering approach to recommend tv programs enhanced with singular value decomposition," *Information Sciences*, vol. 180, no. 22, pp. 4290–4311, 2010.

[11] X. Cai, M. Bain, A. Krzywicki, W. Wobcke, Y. S. Kim, P. Compton, and A. Mahidadia, "Collaborative filtering for people to people recommendation in social networks," in AI 2010: Advances in Artificial Intelligence. Springer, pp. 476–485, 2011.

[12] P. Lops, M. de Gemmis, and G. Semeraro, "Content-based recommender systems: State of the art and trends," in Recommender Systems Handbook. Springer, pp. 73–105, 2011.

[13] W. Carrer-Neto, M. L. Hernández-Alcaraz, R. Valencia-García, and F. García-Sánchez, "Social knowledge-based recommender system. application to the movies domain," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10 990–11 000, 2012.

[14] M. H. Dunham, Data mining: Introductory and advanced topics. Pearson Education India, 2006.

[15] Y. Zhang and E. Cheng, "An optimized method for selection of the initial centers of k-means clustering," in Integrated Uncertainty in Knowledge Modelling and Decision Making. Springer, pp. 149–156, 2013.

[16] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 208–220, 2013.

[17] D. Werner and C. Cruz, "A method to manage the precision difference between items and profiles: In a context of content-based recommender system and vector space model," in *Signal-Image Technology & Internet-Based Systems (SITIS), 2013 International Conference on*. IEEE, pp. 337–344, 2013.

[18] X. Huang and Q. Wu, "Micro-blog commercial word extraction based on improved tf-idf algorithm," in *TENCON 2013-2013 IEEE Region 10 Conference (31194)*. IEEE, pp. 1–5, 2013.