Content-bootstrapped Collaborative Filtering for Medical Article Recommendations

Wenbin Zhang and Jianwu Wang
University of Maryland, Baltimore County, MD, USA
{wenbinzhang, jianwu}@umbc.edu

***Abstract*— Recommender system seeks to assist and augment the natural social process of making choices without sufficient personal experience of the alternatives. They have become fundamental applications in electronic commerce and information access, assisting users to effectively pinpoint information that of their interests from large catalog spaces. Contrary to the pervasive utilization of recommender systems in domains such as electronic commerce, the application of recommendation system in medical domain is limited and further effort is needed. In addition, while a variety of approaches have been proposed for performing recommendation, including collaborative filtering, demographic recommender and other techniques, each individual method has its own drawbacks. This paper proposes a medical oriented recommendation system in which patient's background data is used to bootstrap the collaborative filtering engine and personalized suggestions are provided therein. We present empirical experiment results that show how the content-bootstrapped part of the system enhances the effectiveness of medical article recommendation of the collaborative filtering.**

***Keywords*—Recommender system, collaborative filtering, content-bootstrapped, medical article recommendation.**

I. Introduction

The continuous development of machine learning in different domains has produced numerous algorithms to form meaningful shapes out of them [28], [26], [16]. Recommender system is one of the widespread applications among these learning algorithms. In everyday life, we depend on suggestions from other people either by word of mouth, music recommendation, movie reviews, or reference letters. Recommender systems have emerged to help with this natural social process and to represent user preferences which seek to suggest items for purchase or examination. Recommender systems have become increasingly popular in recent years with a great number of applications in different fields, such as music, movies, and products in general, of which predominantly in the domain of electronic commerce [21], [4]. In a typical recommender system, the main component, recommendation algorithm aggregates the other two components, background data and input data, to arrive its suggestions and directs to appropriate recipients. Background data refers to the information that the system has prior to the beginning of recommendation process, and input data is the profile and behavior information that user communicates to the system in order to generate a recommendation.

A number of recommendation techniques can be distinguished based on the difference between these three components. Among them, one prevalent approach to build recommender system coined the phrase "collaborative filtering" [23]. It aggregates data, such as ratings or recommendations of objects, about users custom or preference, then generates new recommendations to other users based on the recognized commonalities between inter-users and continuously augmented as the user interacts with the system over time. Take the cuckoo movie recommender system [12] as an example, users who shared movie tastes or had expressed mutual cinematic identities could get recommendations from other groups about movies that others with alike taste or identity liked as well. That is to say, the user will be recommended items that people with similar tastes and preferences liked. The greatest strength of collaborative techniques is that they can be content independent, i.e., they can work well in the domains where the availability of content that associated with ratings or recommendations of objects is limited, and the chance of a good match increases with the systems augment by interacting with users continuously. However, this beneficial property of collaborative filtering systems is also its downside. A collaborative filtering technique suffers from a very high item-to-user ratio and lack of first-rater problem in the initial stage of use, it is difficult and unlikely for a system to make quite wisdom suggestions with a small base of rankings or recommendation of objects [8].

Different from collaborative filtering recommendation, content-based systems [17] are absorbed in their associated features and generate recommendations by comparing representations of content therein to representations of content that of user's interests. Content-based recommendation is an outgrowth of information retrieval and filtering research [3], due to the varying degree of early success achieved by the information retrieval and filtering community, many current content-based recommendation systems concentrate on recommending items containing textual information. Content-based recommendation improves the traditional information filtering approaches by incorporating user profile into the recommending procedure. The profiling information shows the user's interests, preferences and needs, and can be elicited from user explicitly, e.g., the description of user preference provided to the system by user directly, or implicitly-extracted by analyzing their transactional behavior over time. The user will be recommended items similar to the ones the user preferred. Specifically, a content-based recommender firstly extracts a set of features from a item, then learns a profile of the user's preferences and determines appropriate items for

978-1-5386-5488-0/18/$31.00 ©2018 IEEE

recommendation purpose. Therefore, content-based techniques are feasible when the system has a small base of rankings or recommendation of objects.

Another category of recommendation techniques is demographic recommender systems [6]. As its name suggests, demographic recommendation aims to provide recommendations based on the demographic classes categorized according to user's personal attributes. For instance, consider a products and services recommender system proposed by Krulwich [13]. The system first uses a short survey to gather user categorization information explicitly, then the formed demographic groups is used for the preparation of products and services recommendation. The general approach of this kind of systems is to match users' responses against a library of assembled user stereotypes. Just as collaborative approaches, demographic methods make recommendation based on the formed "people-to-people" correlations, but use different data. The user will be recommended items that people belongs to the same demographic class liked. A demographic technique does not require a history of user's ratings but is hard to reflect the changes of a user's interests over time.

In addition, knowledge based recommendation [24] has knowledge about how a particular item meets a particular user's interests, and can therefore make individualized suggestions by reasoning about user's needs and preferences. The benefit of a knowledge based approach is that it is capable of generating recommendation regardless the availability of user's rating. However, the building of knowledge engineering is labor-intensive and time consuming. Utility based recommender [11] attempts to suggest object based on a computation of the utility of each object of the user. The profile of each individual user therefore is the derived user-specific utility function and the system applies it to the objects under consideration for the best match. The utility function is able to incorporate non-product attributes, such as vendor reputation and delivery service, into the utility computation, making for instance the user with immediate need will be recommended items that trades off price against delivery schedule. [15], [18], [2] give excellent reviews and heaps of examples of recommendation systems.

Including the aforementioned recommender techniques, a variety of recommender approaches have been proposed by research community with applications in numerous fields [2], [14], [22]. They are now an integral part of some on-line recommendation services providing stores such as Amazon and CDNow [20]. Contrary to the wide and predominant utilization in the electronic commerce area, the application of recommendation system in other domains has been underexplored [5], [25]. The use of recommendation systems in alternative areas such as medical domain is also promising and a more detailed study is warranted. With this motivation in mind, this paper describes a collaborative filtering recommendation system with an application in medical domain. To alleviate the high item-to-user ratio and lack of first-rater drawbacks, a content-based predictor is used to bootstrap the collaborative filtering process, with the purpose of providing an enhanced personalized suggestion.

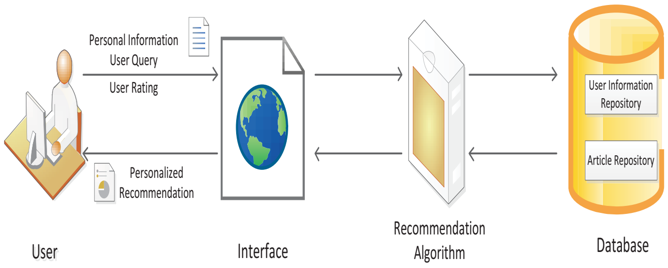

Fig. 1: The architecture of proposed recommender system.

The remainder of this paper is organized as follows. In section II, we propose the medical domain oriented content-bootstrapped collaborative filtering recommendation system, and in section III we perform empirical experiment to compare our proposed approach with other methods and analyze the obtained results. Section IV concludes this work and discusses possible further research.

II. The content-bootstrapped collaborative filtering

Our proposed system works in the domain of medical recommendation, which provides medical article recommendations to patients confronting lung cancer or lung cancer practitioners. The system consists three main components. The standard recommendation process of the system starts with a typical user interactively communicating with the system through the first component, namely user interface. The user profile information, which comprises common information needs of lung cancer patients [19], being used as background data is required to be filled in for the first-time user or each time a practitioner has a particular need. User also selects one of the four target categories of lung cancer articles, which are risk factors, life style, emotional concerns and treatment (each category associates with fixed number of features), provided by previous lung cancer article classification studies [1], and rates the recommended article on a 0-5 integral scale as the input data. These two data are then utilized by the second component, recommendation algorithm, to know users and in return facilitate users' effort in finding articles meet with their satisfactions. The third database component stores background data and input data in user information repository and different topics of lung cancer articles in article repository. The general overview of the system is shown in Figure 1. In the following, we describe the implementation of the content-based predictor and its bootstrapped collaborative filtering recommendation in details, which is the core part of the system and distinguishes our system from others. Table I gives a summary of notations used for problem definition and method description.

In the domains like electronic commerce where collaborative filtering recommender systems have been predominantly

Notation	Description
a	Target user, of whom recommendation is being made
b	User other than target user
c	Target article
d	Article other than target article
α	Significance weighting factor
USim	User similarity between users
ASim	Article similarity between articles
USet	User set in the system
NSet	Formed neighborhood of the target user
ASet	Articles from the same group of target article

TABLE I: Notation used for problem definition and method description.

utilized, suggestions are made to other users based on similarity between users in overall patterns. Such commonalities between inter-users are commonly recognized by comparing users' rating on the co-rated items. The intuition of application in such domain is that users who previously shared taste or had co-expressed their interests in certain items are more prone to show similar preference in other items. However, in medical article recommendation domain, this is counter-intuitive. For example, patients shared similar risk factors in the past deviate their particular article interests after their risk factors become different. It means the previous similarity of needs reflects from the co-rated articles might not be useful to determine the neighborhood users with similar current demands. Considering this unique character of recommender application in medical article domain, our intuition is to follow the proverb - "birds of a feather flock together" to overcome this difficulty. Therefore, in contrast with the neighboorhood forming process of other recommender applications, we determine to exclusively choose the target user's neighborhood based on the user information. Specifically, the shared features of four categories along with demographic information are used to compute the similarity between two users, and the user similarity (*USim*) between user *a* (for whom the recommendation are being made) and user *b* is defined as follows:

$$USim = \frac{\sum_{i=1}^{n} N_{i,a\cap b}}{\sum_{i=1}^{n} N_i} \tag{1}$$

where N_i is the number of features of each group, $N_{i,a\cap b}$ corresponds to the number of shared features between user *a* and user *b* in group *i* and *n* is the number of categories of the article along with user's background data (*n* equals to 5 in this particular application). However, we may want to give one of the four categroy groups that a user requests recommendation more importance than the remaining four groups. In other words, we would like to increase the confidence we place in the group that of user *a*'s particular interest. We do this by multiplying both the number of features of the requested group and the number of shared features between users in this group by two when computing N_i and $N_{i,a\cap b}$. In addition, it is natural for users to share certain commonly overlapped features such as gender and marital status. We would like to devalue the user similarity based on these features and consequently revalue user similarity with comparatively more shared features. To align with this idea, we multiply the *USim* by a significance weighting factor [7]. In particular, when the two users have less than half of the total number of features from all groups in common, *USim* is multiply by a factor $\alpha = n/N$, where *n* represents the number of shared features and *N* is the total number of features of all groups, and *USim* stays unchanged otherwise, i.e. $\alpha = 1$.

$$USim = \begin{cases} \alpha \cdot USim & if\ n\ <\ N/2 \\ USim & if\ n\ \geq\ N/2 \end{cases} \tag{2}$$

Thus, user similarity will be devalued appropriately if two users have less than half of the total number of features of all groups, and the weighting of similarity between other users will also be revalued correspondingly. In our evaluation, following the recommendation of [7], top 20 users that have the highest similarity with the target user are selected to form the neighborhood of the target user, i.e., user *a*.

With the selected neighbors of user *a*, the prediction of target user rates a target article *c* is dependent on the availability of rating provided by the neighborhood of user *a* to article *c*. As discussed in section I, collaborative filtering techniques suffer from a very high item-to-user ratio and lack of first-rater problem in the initial stage of use. We overcome these drawbacks of collaborative filtering approaches by exploiting similar articles from the same group already rated by neighbors of the target user. Specifically, we first view articles as a "bag" of words or terms and each article is represented as a vector, then each term weight is computed based on the *TF-IDF* term weighting scheme [10], which is given by the following formula:

$$w_{m,i} = \frac{f_{mi}}{\max\left\{f_{1i}, f_{2i}, \cdots, f_{|v|i}\right\}} \times \log\frac{N}{df_m} \tag{3}$$

where f_{mi} is the number of times that term t_m appears in article d_i, *N* represents the total number of articles, and df_m corresponds to the number of articles that t_m appears. Afterwards, article similarity (*ASim*) between article *c* and article *d* is computed using the cosine similarity coefficient [10], defined below:

$$ASim(d_c, d_d) = \frac{\sum w_{m,c} \cdot w_{m,d}}{\sqrt{\sum w_{m,c}^2 \cdot w_{m,d}^2}} \tag{4}$$

The learned article similarity is then used to provide ratings for neighbors of user *a* to article *c*, where the availability of rating is lacked as none of them has rated the target article before. Denote the user-rating vector of article *c* rated by user *a*'s neighbors as $v_{c,b\in neighbor\ of\ a}$, this sparse vector is convert to a dense one as follows:

$$v_{c,b\in neighbor\ of\ a} = \begin{cases} v_{c,b\in neighbor\ of\ a} & if\ v_{c,b\in neighbor\ of\ a} \neq \emptyset \\ \frac{\sum_{d\in D_i} ASim(d_c,d_d)\cdot r_{d,b}}{\sum_{d\in D_i} ASim(d_c,d_d)} & if\ v_{c,b\in neighbor\ of\ a} = \emptyset \end{cases} \tag{5}$$

where D_i is the set of articles already rated by neighbors of the target user a and from the same group of target article d_c. We now perform collaborative filtering using the densified vectors and the final prediction for target article c rated by target user a is defined below:

$$r_{c,a} = \frac{\sum_{b\in neighbor\ of\ a} USim(a,b)\cdot r_{c,b}}{\sum_{b\in neighbor\ of\ a} USim(a,b)} \tag{6}$$

where $r_{c,b}$ is the actual rating provided by one of the target user a's neighbors, user b, to the target article c, where available, otherwise is the predicted rating by user b to article c when none of the selected neighbors has rated the target article before. Next, the top 10 rated articles will be recommended to the target user. The user rates how these articles meet his/her information needs and the rating information is stored in the database to augment the process of system recommendation. The main procedure of the proposed content-bootstrapped collaborative filtering is detailed in Algorithm 1 with notations that are identical with representations used for problem definition and method description in the previous part of the paper.

Algorithm 1: The content-boostrapped collaborative filtering algorithm

```
Input: One of the four target categories
Output: Recommended articles
while USet has more users do
    Calculate USim according to Equation (1) and (2);
    Forming NSet;
end
for user $b \in NSet$ do
    if $\exists$ user $b$ rates $d_c$ then
        Calculate rating by user $a$ to $d_c$ according to Equation (6);
    else
        for $d_d \in ASet$ do
            Calculate ASim according to Equation (4);
            Calculate $v_{c,b\in neighbor\ of\ a}$ according to Equation (5);
            Calculate rating by user $a$ to $d_c$ according to Equation (6);
        end
    end
end
Recommend top ranked articles.
```

III. Experiments

This section presents an empirical study designed to evaluate the performance of our proposed content-bootstrapped collaborative filtering in the domain of medical article recommendation. We selected 20 articles each from the abovementioned four categories of recommendations and stored them in the article repository to cover the common needs of information accessing of lung cancer patients or lung cancer practitioners. In addition, 9 patients' background data was used with consent and ethics clearance, we rated articles on their behalf and further simulated additional 91 patients' background data as well as input data. A group of 3 students working on the expression status of biomarkers in lung cancer along with 11 graduate students taking biostatistics course were acted as the lung cancer practitioners in evaluating how the information needs of lung cancer practitioners is met by our proposed method (*CBCF*) against aforementioned collaborative filtering based recommendation [23] and content-based systems [17]. The evaluation is measured by two widely used evaluation metrics for recommenders, mean absolute error (MAE) and Receiver Operating Characteristics (ROC) [9]. MAE metric is a statistical accuracy metric defined as the average absolute difference between predicted ratings and the users' true ratings. In other words, it shows the deviation between rating predicted by recommender system and actual rating given by user. The other metric ROC measures how well an information system can distinguish between relevant and non-relevant items. In our experiments, we classify an article as highly relevant when user rates this article with a rating higher than 4 and thus accept this article, referred as sensitivity, otherwise we consider it of low relevance with rejection, referred as specificity. The value of ROC sensitivity ranges from 0 to 1 with a higher value indicates a better prediction for high-quality article. Table II shows the summarized results of all methods w.r.t. MAE and ROC metrics.

Method	MAE	ROC
Collaborative filtering	1.002	0.587
Content-based	1.123	0.632
CBCF	**0.912**	**0.683**

TABLE II: Performance comparison w.r.t MAE and ROC metrics.

From Table II, it is clear that the proposed method (the bottom row) beats all other methods on both metrics. In particular, the proposed method performs 18.8% better than content-based system and 9% better than collaborative filtering based recommendation on the MAE metric. This depicts the proposed approach, compared to other methods, does a better job of recommending articles that meet the needs of the users by providing closer to actual rating prediction, and is a reliable predictor of article recommendation. In this regard, the contributions of the superior selection of neighborhood and more representative neighborhood similarity calculation to the improvement of recommendation of the proposed method in medical domain are verified. In terms of ROC metric, the proposed method outperformances content-based system and collaborative filtering based recommender by 8.1% and 16.4%, respectively. This result agrees with the more accurate prediction provided by the proposed approach on MAE metric and demonstrates the capability of recommending high-quality articles of the method.

IV. Conclusions

A variety of recommender approaches have been proposed by research community and have been widely applied in electronic commerce. On the contrary, the application of recommendation system in other domains has been under-explored. This paper describes a collaborative filtering recommendation system with an application in the alternative medical domain. To solve the high item-to-user ratio and lack of first-rater problem and produce more reliable predictions thereafter, a content-based predictor is used to bootstrap the collaborative filtering process. In addition, considering the unique characteristics of information needs of medical domain, the selection of target user's neighborhood is based on the user information exclusively. The results of empirical study suggest the proposed approach is an effective medical article recommender. An online system will be deployed based on our previous streaming framework [27] to provide real-time recommendations in the future. It is anticipated that further exploration and research in this promising domain is to come.

References

[1] MedlinePlus. https://medlineplus.gov/.

[2] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6):734–749, 2005.

[3] N. J. Belkin and W. B. Croft. Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, 35(12):29–38, 1992.

[4] J. Boehmer, Y. Jung, and R. Wash. e-commerce recommender systems. *The International Encyclopedia of Digital Communication and Society*, 2015.

[5] 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*, 5(2):169–181, 2011.

[6] M. A. Ghazanfar and A. Prugel-Bennett. A scalable, accurate hybrid recommender system. In *Knowledge Discovery and Data Mining, 2010. WKDD'10. Third International Conference on*, pages 94–98. IEEE, 2010.

[7] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 230–237. ACM, 1999.

[8] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[9] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[10] A. Huang. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, pages 49–56, 2008.

[11] S.-L. Huang. Designing utility-based recommender systems for e-commerce: Evaluation of preference-elicitation methods. *Electronic Commerce Research and Applications*, 10(4):398–407, 2011.

[12] R. Katarya and O. P. Verma. An effective collaborative movie recommender system with cuckoo search. *Egyptian Informatics Journal*, 2016.

[13] B. Krulwich. Lifestyle finder: Intelligent user profiling using large-scale demographic data. *AI magazine*, 18(2):37, 1997.

[14] S. K. Lam and J. Riedl. Shilling recommender systems for fun and profit. In *Proceedings of the 13th international conference on World Wide Web*, pages 393–402. ACM, 2004.

[15] P. Lops, M. De Gemmis, and G. Semeraro. Content-based recommender systems: State of the art and trends. In *Recommender systems handbook*, pages 73–105. Springer, 2011.

[16] K. Mo, T. Zhao, W. Zhang, and S. Wang. Quality assessment of timber forest at sub-compartment level: the algorithm and its accuracy. *indicator*, 17:18, 2016.

[17] R. J. Mooney and L. Roy. Content-based book recommending using learning for text categorization. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 195–204. ACM, 2000.

[18] F. Ricci, L. Rokach, and B. Shapira. Introduction to recommender systems handbook. In *Recommender systems handbook*, pages 1–35. Springer, 2011.

[19] R. Sanson-Fisher, A. Girgis, A. Boyes, B. Bonevski, L. Burton, and P. Cook. The unmet supportive care needs of patients with cancer. *Cancer*, 88(1):226–237, 2000.

[20] J. B. Schafer, J. Konstan, and J. Riedl. Recommender systems in e-commerce. In *Proceedings of the 1st ACM conference on Electronic commerce*, pages 158–166. ACM, 1999.

[21] J. B. Schafer, J. A. Konstan, and J. Riedl. E-commerce recommendation applications. pages 115–153, 2001.

[22] J. B. Schafer, J. A. Konstan, and J. Riedl. E-commerce recommendation applications. In *Applications of data mining to electronic commerce*, pages 115–153. Springer, 2001.

[23] U. Shardanand and P. Maes. Social information filtering: algorithms for automating word of mouth. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 210–217. ACM Press/Addison-Wesley Publishing Co., 1995.

[24] S. Trewin. Knowledge-based recommender systems. *Encyclopedia of library and information science*, 69(Supplement 32):180, 2000.

[25] L. Zhang and W. Zhang. A comparison of different pattern recognition methods with entropy based feature reduction in early breast cancer classification. *European Scientific Journal, ESJ*, 10(7), 2014.

[26] W. Zhang, J. Tang, and N. Wang. Using the machine learning approach to predict patient survival from high-dimensional survival data. In *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*, pages 1234–1238. IEEE, 2016.

[27] W. Zhang and J. Wang. A hybrid learning framework for imbalanced stream classification. In *Big Data (BigData Congress), 2017 IEEE International Congress on*, pages 480–487. IEEE, 2017.

[28] W. Zhang, J. Wang, D. Jin, L. Oreopoulos, and Z. Zhang. A deterministic self-organizing map approach and its application on satellite data based cloud type classification. In *Big Data (Big Data), 2018 IEEE International Conference on*, pages 1204–1211. IEEE, 2018.