Automatically Recommending Healthy Living Programs to Patients with Chronic Diseases through Hybrid Content-Based and Collaborative Filtering

Yizhou Zang, Yuan An, Xiaohua Tony Hu
College of Computing and Informatics, Drexel University
Philadelphia, USA
{yz388, ya45, xh29}@drexel.edu

***Abstract*—In this research, we develop a hybrid recommendation system recommendation system for healthy living programs to patients with chronic diseases. Our experiments indicate that our model compared favorably against other real-world recommendation applications in terms of accuracy. We also demonstrated that the proposed hybrid algorithm performed better than traditional CF in terms of error rate, precision and recall.**

Keywords—collaborative filtering, hybrid, EMR, Healthy Living Program, chronic dieseases

I. INTRODUCTION

Chronic diseases such as diabetes and hypertension are among the most preventable health problems. Strong evidence has shown that preventive approaches such as participating in healthy living programs can achieve good results for controlling chronic diseases. As health IT prevails, large amounts of longitudinal data about patient health status are accumulated in various electronic health record (EHR) systems. The data provide a great opportunity to explore evidence-based methods for helping patients control chronic diseases.

In this research, we develop and evaluate an automated recommendation system for healthy living programs to patients with chronic diseases. The recommendation system takes as input patient clinical data such as health conditions and vital observations and wellness data such as program attendance and health screening survey results. It then recommends a set of healthy living programs to a patient by a hybrid technique combining both content-based and collaborative filtering methods based on the evidence gathered from the patients with similar conditions and outcomes (ratings).

We are motivated by the problem of using health information technology to improve the healthcare outcomes at a comprehensive nurse-managed community health center. The center provides a wide variety of healthy living programs and wellness services including physical exams, diagnosis and treatment of illness, family planning, health maintenance/disease prevention services, behavioral health services, physical fitness programs, dental services, nutrition services, and chronic disease management programs.

Through the years, the center has been striving to serve more patients and has tracked various patient information using health information technologies. Specifically, the center has developed and implemented a comprehensive health information infrastructure [1] consisting of an Electronic Medical Record (EMR) and a Patient Wellness Tracking (PWT) system. The EMR documents patients encounter information, medical history, medications, and lab test results, while the PWT, linking to the EMR, contains data on behavioral screening, health assessment, healthy living programs (HLP) and social activities.

Through this research we introduce the hybrid recommender system as an important function of evidence-based healthcare to the comprehensive health information system at the center. The main goal is to assist clinicians at the center to assign patients to healthy living programs based on the evidence gleaned from the large amount of longitudinal data.

II. RELATED WORK

Data mining techniques are becoming more and more popular in healthcare domain because the data generated by healthcare organizations is too complex and vast to be processed and analyzed by traditional methods. Data mining fetches unknown patterns and useful information from huge data sets, helping healthcare specialists make medical decisions such as estimation of medical staff, health insurance policy formulation, disease protection, treatment selection, etc. [2,3].

However, there is a lack of study on exploiting recommender system techniques for healthcare decision-making. Although Kahn et al. 2005 and Davis et al. 2009 discuss the use of collaborative filtering in healthcare, there is no concrete study on a specific healthcare problem [4][5].

III. THE DESIGN OF THE RECOMMENDER MODULE

We use the dataset from a nurse-managed health service center. The dataset comprises of two parts: (1) encounter information, medical history, medications, and lab test results of 5724 patients documented in EMR, and (2) information about 477 Healthy Living Programs and corresponding wellness data such as program attendance and health screening survey results collected in PWT. The module consists of 3 components: Component A pre-processes the data. Component B is responsible for mapping patients' clinical and wellness data to a rating system. Component C is responsible for making recommendation through a hybrid recommendation approach.

978-1-4799-5669-2/14/$31.00 ©2014 IEEE

The Details of each individual component are discussed in the following sub-sections.

A. Preprocessing Data

A traditional collaborative filtering recommender system predicts target user's rating for the target item, based on the users' ratings on observed items. So the user-item rating matrix is the key. In our health information system, we utilize patients' health survey records instead. Patients are periodically asked by center clinicians to take several health-screening surveys. We believe that the health survey scores reflect patient's health conditions at a certain point. Consequently, we believe that the change of survey scores at two time points can reflect the change of patient's health conditions between these two time points.

Furthermore, if a patient took a healthy living program (HLP) during this period of time, assuming that this patient's health condition changes are mainly due to the participation in the HLP, we can state that the change of survey scores indicate whether or not the HLP works. Thus, we can map the change of survey scores into normalized score R_e representing how useful the target health living program is for the target patient. The assumption that patient's health condition changes are due to the participation in HLPs is reasonable, because we use all the other clinical and treatment conditions for measuring the similarity between patients. Although in this paper we restrict ourselves to several important vital signs as the clinical conditions, we believe the overall approach can be extended straightforwardly to incorporate all the conditions.

Because of all of these, patients who do not have records of health screening surveys will be excluded.

B. Mapping survey data to a rating system

1) Selecting proper health surveys

In PWT system, there are in total 37 different surveys provided at present. Among these surveys, 4 health surveys are selected in our module: SF-36 health survey[1], PHQ-9 survey[2]: Nine Symptom Checklist, GAD-7 survey, and PHQ-9 and GAD-7 Screening survey[3].

2) Identifing the time period during which the changes of score occurred

In our dataset, there are a few cases where a patient takes several HLPs at the same time, or takes one HLP several times. The exceptions have been removed in the pre-processing. Therefore, for each (patient, program) pair, we pick the latest time point before the target patient participates in the target health living program, and the earliest time point after the patient finishes the target program.

3) Transferring survey changes into a rating on a 5–point scale

For each (patient, program) pair, we've already got two sets of surveys taken respectively at two time points. In this step, we calculate the score change of each survey, and then normalized the change into a 5–point scale. At last, we combine the four normalized scores ($R_{sf36}, R_{phq9}, R_{gad7}$, R_{p9g7}) into a final rating R_e.

Since different surveys have different scoring mechanisms, we convert the absolute score changes of the 4 surveys into a unified 5-point scale based on different rules as follows.

For a SF-36 survey, there are two major scores - Physical Component Summary (PCS) and Mental Component Summary (MCS). We convert these two scores separately as shown in Table 1. PCS and MCS scores (R_{PCS}, R_{MCS}) are equally weighted when calculating the overall SF-36 rating R_{sf36} as shown in Eq 1.

Score change of PCS	PCS Rating (R_{pcs})	Score change of MCS	MCS Rating (R_{mcs})
< -20	1	< -20	1
-20 ~-10	2	-20 ~-5	2
-10 ~ 0	3	-5 ~ 10	3
0 ~ 10	4	10 ~ 25	4
> 10	5	> 25	5

Table1. SF36 Converting Table

$$R_{sf36} = (\frac{1}{2}R_{PCS} + \frac{1}{2}R_{MCS})/2 \quad (1)$$

For PHQ9 survey, the converting standard we use is shown in Table 2.

Score change of PHQ9	PHQ9 Rating (R_{phq9})
< -5	1
-5 ~ 0	2
0 ~ 5	3
5 ~ 10	4
> 10	5

Table2. PHQ9 Converting Table

For GAD7 survey, the converting standard we use is shown in Table 3.

Score change of GAD7	GAD7Rating (R_{gad7})
< -20	1
-20 ~ -10	2
-10 ~ 0	3
0 ~ 10	4
> 10	5

Table3. GAD7 Converting Table

For PHQ-9 and GAD-7 Screening survey, the converting standard we use is shown in Table 4.

Score change of PHQ-9 and GAD-7 Screening	PHQ-9 and GAD-7 Screening Rating (R_{p9g7})
< -20	1
-20 ~ -10	2
-10 ~ 0	3
0 ~ 10	4
> 10	5

Table4. PHQ9 and GAD7 Screening Converting Table

We set the above converting standards based on the distribution of score change. For example, since the score changes for PCS range are from -24.39 to 15.34, the

[1] http://www.sf-36.org/tools/sf36.shtml
[2] http://www.phqscreeners.com/
[3] http://www.phqscreeners.com/

converting standard we set in Table 1 covers this range and leaves some room for future data that might exceed this range. Then we evenly divide this range into 5 internals and assign a corresponding rating to each internal.

At each time point, a patient may take all or only some of these four health surveys. Therefore, we combine these four ratings into a single rating as showing:

$$R_e = (R_{sf36} + R_{phq9} + R_{gad7} + R_{p9q7})/N_s \qquad (2)$$

In Eq. 2, N_s is the number of surveys the target patient takes during a certain period of time. The four scores of SF-36, PHQ9, GAD7 and PHQ-9 and GAD-7 Screening are equally weighted when calculating the final score (R_e).

4) Building up Patient-Program Matrix

From previous section, for each (patient, program) pair, we already get the corresponding rating, R_e. Therefore, a patient-program matrix then can be created as following:

	Pro-1	Pro-2	…	Pro-n
Pat-1	R_{e11}	R_{e12}		R_{e1n}
Pat-2	R_{e21}	R_{e22}		R_{e2n}
…				
Pat-m	R_{em1}	R_{em2}		R_{emn}

Table5. Patient-Program Matrix

Where each row i represents a patient, i.e., Pat-i, m is the number of patients; each column j represents a HLP, i.e., Pro-j, n is the number of programs.

C. Hybrid recommendation approach

We select representative patient attributes, normalize them into a score on a 5-point scale, and build a patient-attribute matrix based on that. Pearson's correlation is chosen for measuring patient attribute similarity based on the Patient-Attribute Matrix. Pearson's correlation is also used to measure patient rating similarity based on the Patient-Program matrix produced in Step 3.*B*. We combine these two similarities as an overall patient similarity. In the end, we run collaborative filtering on Patient-Program matrix.

1) Selecting patient attributes

Blood pressure, BMI and Hga1c are chosen because they are representative patient's attributes and can be easily measured. In the future, we plan to extend the approach to include a complete set of patient attributes.

2) Converting patient attributes to a score on a 5-point scale

We uniformly select the latest record taken before the target patient took any health living programs. We transfer the value for each attribute into a score on a 5-point scale. The translation is based on common knowledge about health status and vital medical conditions, for example, Blood Pressure Chart[4]. The standards are shown as follows:

Systolic		Diastolic	Category	Rating (R_{bp})
<120	And	<80	Normal	5
120-139	Or	80-89	Prehypertension	4
140-159	Or	90-99	Stage 1 hypertension	3
160-180	Or	100-110	Stage 2 hypertension	2
180+	Or	110+	Stage3&4 hypertension	1

Table6. Blood Pressure Converting Table

BMI	Weight status	Rating (R_{bmi})
18.5 - 24.9	Normal	5
25.0 - 29.9 OR <18.5	Overweight or Underweight	4
30.0 - 34.9	Obese	3
34.9 - 39.9	Severely Obese	2
>40.0	Very Severely Obese	1

Table7. BMI Converting Table

Hga1c	Rating (R_{hga1c})
4.5 - 5.7	5
5.7 - 6.4	4
<4.5	3
6.4 - 8	2
>8	1

Table8. Hga1c Converting Table

3) Building up the Patient-Attribute Matrix

For each patient, we already get the corresponding rating: R_{bp}, R_{bmi} and R_{hga1c}, according to the Section 3.*C*.2. Therefore, a patient-program matrix then can be created as following:

	Blood Pressure	BMI	Hga1c
Pat-1	R_{bp1}	R_{bmi1}	R_{hga1c1}
Pat-2	R_{bp2}	R_{bmi2}	R_{hga1c2}
…			
Pat-m	R_{bpm}	R_{bmim}	R_{hga1cm}

Table9. Patient-Attribute Matrix

Where each row represents a patient, m is the number of patients; each column represents an attribute.

4) Measuring similarity

Pearson's correlation measures the linear correlation between two vectors of rating. We apply Pearson's correlation to measure both patient rating similarity and patient attribute similarity.

The rating similarities between patients can be calculated based on the patient-program matrix generated in Section.3.*B*.4 and equation (3).

$$sim_p(i,j) = \frac{\sum_{p \epsilon I_{ij}}(R_{i,p} - A_i)(R_{j,p} - A_j)}{\sqrt{\sum_{p \epsilon I_{ij}}(R_{i,p} - A_i)^2 \sum_{p \epsilon I_{ij}}(R_{j,p} - A_j)^2}} \qquad (3)$$

Where $R_{i,p}$ is the rating of the program p by patient i, A_i is the average score of user i for all the co-scored programs, and I_{ij} is the program set both scored by patient i and patient j.

Similarly, the attribute similarities between patients can be calculated based on the patient-attribute matrix generated in Section.3.*C*.3 and equation (4).

[4]http://www.heart.org/HEARTORG/Conditions/HighBloodPressure/AboutHighBloodPressure/Understanding-Blood-Pressure-Readings_UCM_301764_Article.jsp

$$sim_a(i,j) = \frac{\sum_{a \epsilon I_{ij}}(R_{i,a} - A_i)(R_{j,a} - A_j)}{\sqrt{\sum_{a \epsilon I_{ij}}(R_{i,a} - A_i)^2 \sum_{a \epsilon I_{ij}}(R_{j,a} - A_j)^2}} \quad (4)$$

Where $R_{i,a}$ is the score of the attribute a by patient i, A_i is the average score of user i for all the co-scored attributes, and I_{ij} is the attribute set both scored by patient i and patient j.

Then, we combine these two similarities into a composite measure as shown in equation (5):

$$sim(i,j) = \omega sim_p(i,j) + (1-\omega) sim_a(i,j) \quad (5)$$

Where $sim_p(i,j)$ is patient rating similarity, $sim_a(i,j)$ is patient attribute similarity, and ω and $1-\omega$ are coefficients determining the importance of each similarity.

5) Choosing Neighbors

According to the idea of CF, the most similar users, whose ratings are used for predicting ratings of target user, are called neighbors. There are two major methods for selection of neighbors: (1) setting a threshold (2) choosing top-N neighbors. In this model, we utilize the top-N approach.

6) Making Recommendation

Once we get the rating matrix, the similarity between patients, we can calculate the predicted rating of target patient to the target program as follows:

$$P_{ui} = A_u + \frac{\sum_{i=1}^{n}(R_{it} - A_i) * sim(u,i)}{\sum_{i=1}^{n} sim(u,i)} \quad (6)$$

Where A_u denotes the average rating of the target patient u to the programs, R_{it} denotes the rating of the neighbor patient i to the target program t, A_i is the average rating of the neighbor patient i to the programs, $sim\,(u,i)$ is the similarity of the target patient u and the neighbor patient i, and n is the number of the neighbors.

IV. EXPERIMENT

A. Performance Measurement

In evaluating the performance of the proposed method, we are concerned with two categories of accuracy metrics: *statistical accuracy metrics* and *decision-support metrics*.

1) Statistical Accuracy Metric

Mean absolute error (MAE) is a statistical accuracy metric to access the accuracy of a prediction algorithm by comparing the numerical deviation of the predicted ratings from the actual ratings, as shown in equation (7):

$$MAE = \frac{\sum_{i=1}^{n}|p_i - q_i|}{n} \quad (7)$$

Where n is the total number of ratings, p_i is the predicted rating and q_i is the real rating. The lower the MAE, more accurate the prediction is.

2) Decision-support metrics

Precision and Recall are employed in this research. Let R denote the number of relevant recommended items, N denote the number of recommended items and U denote the number of all relevant items. Then, Precision is the radio of R to N, while Recall is the radio of R to U.

B. Dataset

After the data pre-processing, 862 ratings from 118 patients on 477 healthy living programs are retained. We divide the 118 patients into 5 groups, 3 of which consists of 24 patients and 2 of which consists of 23 patients. For each group, we regard it as test set, while the rest of the 4 groups are regarded as training set. We conduct cross-validation by repeating the experiment 5 times and each time a different group is regarded as a test set. At last, an average of these 5 experiments is calculated.

C. Results

1) Determining Coeffecient ω

We first tested the change of MAE of proposed method with the coefficient ω. ω was set from 0.1 to 0.9 with an increment of 0.1. We examine the MAE- ω curve when N (number of neighbors) takes different values. The MAE values generated were shown in Fig2.

We can observe that all the MAE values are between 0.5 and 1.2. And when top-30 nearest neighbors are selected, MAE values are reduced to around 0.6. Although different recommender systems vary significantly, for one with a 5-point rating system, such MAE values are reasonable. According to studies on empirical performance of real-world recommendation applications, such as the famous MovieLens and Netflix, the MAE values are usually between 0.5 and 0.8. This indicates that our module compares favorably against other real-world recommender systems.

From the observation, we can also find that the lowest MAE values obtained when ω is around 0.3. Therefore, we use 0.3 as the optimal value for coefficient ω, and compare our method with traditional CF in terms of MAE, Precision and Recall in next section.

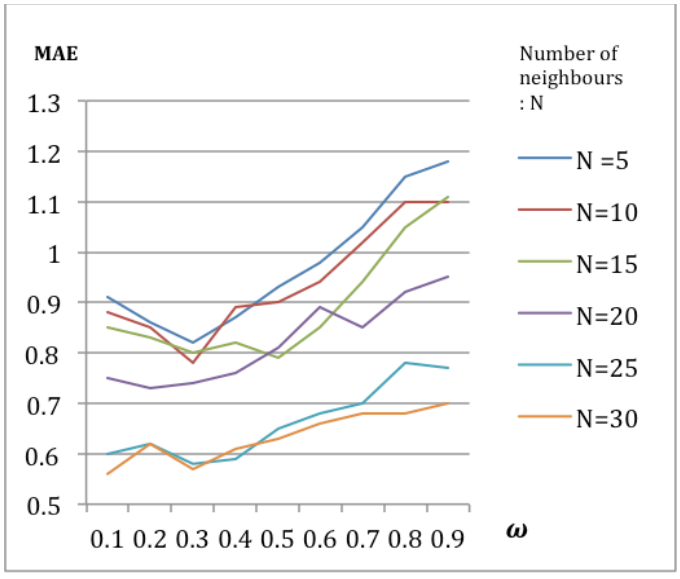

Fig.2 MAE with respect to the similarity combination coefficient ω

2) Comparing with traditional CF

In this section, we compare our hybrid algorithm with the traditional use-based CF with respect to MAE, precision and recall.

Fig. 3 demonstrates MAE of proposed method and traditional CF. The obvious conclusion is that our method, with lower MAE value, consistently performs better than traditional CF.

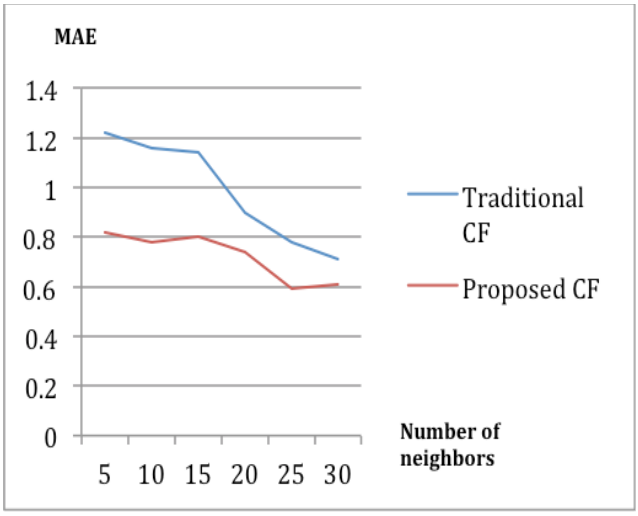

Fig. 3 Comparison of MAE between proposed algorithm and traditional CF

We also examined the decision-support metrics of proposed method and traditional CF. According to section 4.A.2, Precision is the radio of R (number of relevant recommended items) to N (number of recommendations), while Recall is the radio of R to U (the number of all relevant items). We define the number of recommended programs to be N, all the HLPs assigned to patients by the clinicians to be useful programs (P). Thus, Precision= (N ∩ P) /N and Recall= (N ∩ P) / P. Then, we examined the change of precision/recall with number of recommendations.

Fig.4 and Fig.5 respectively show the precision and recall in relation to the number of recommendations. In both figures, our proposed CF has higher precision or recall value than traditional CF, demonstrating that our method has better performance of simulating clinician's diagnose.

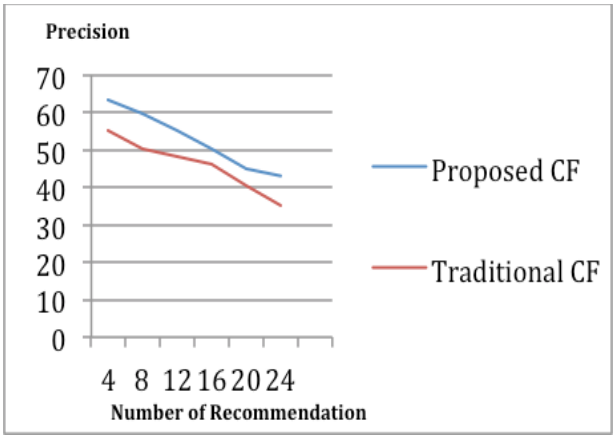

Fig.4 Comparison of Precision between proposed algorithm and traditional CF

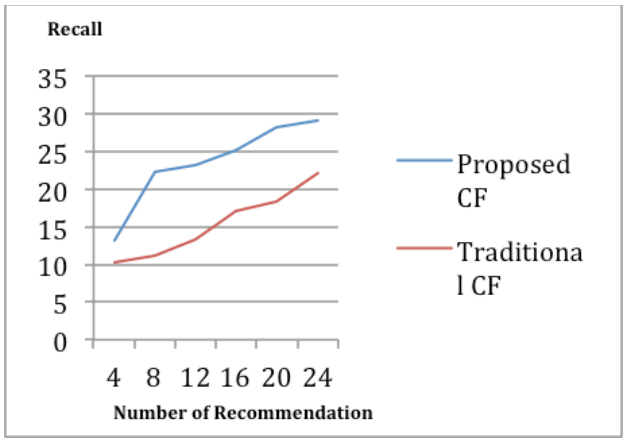

Fig.5 Comparison of Recall between proposed algorithm and traditional CF

V. CONCLUSION

In this research, we proposed a recommendation model that automates the process of assigning healthy living programs to patients at a health services center. We constructed a patient-program rating matrix and patient-attribute matrix based on the patients' HLP data and clinical data stored in our PWT system and EMR system respectively. We then developed a hybrid recommendation approach, which includes both content-based and collaborative filtering recommendation methods.

Our experiments indicated that our model compared favorably against other real-world recommendation applications in terms of prediction quality. We also proved that the proposed hybrid algorithm performed better than traditional CF in terms of MAE, precision and recall.

As for follow-up researches, we will focus on two aspects. (1) We plan to evaluate the system through quantitative study on real patients and providers. (2) We will improve our module by taking into account more patients' attributes and health surveys.

ACKNOWLEDGMENT

This work is supported in part by a Drexel Jumpstart grant on Health Informatics and the NSF grant IIP 1160960 for the center for visual and decision informatics (CVDI).

REFERENCES

[1] An, Y., Dalrymple, P.W., Rogers, M., Gerrity, P., Horkoff, J., Yu, E.: Collaborative Social Modeling for Designing a Patient Wellness Tracking System in a Nurse-Managed Health Care Center. 4th Int. Conf. on Design Science Research in Information Systems and Technology (DESRIST) (2009)

[2] M. Silver, T. Sakara, H. C. Su, C. Herman, S. B. Dolins and M. J. O'shea, "Case study: how to apply data mining techniques in a healthcare data warehouse", Healthc. Inf. Manage, vol. 15, no. 2, (2001), pp. 155-164.

[3] V. S. Stel, S. M. Pluijm, D. J. Deeg, J. H. Smit, L. M. Bouter and P. Lips, "A classification tree for predicting recurrent falling in community-dwelling older persons", J. Am. Geriatr. Soc., vol. 51, (2003), pp. 1356-1364

[4] Kahn CE Jr (2005) Collaborative filtering to improve navigation of large radiology knowledge resources. J Digit Imaging 18(2):131–137

[5] Davis, D. A., Chawla, N. V., Christakis, N. A. and Barabsi, A.-L. (2009). Time to CARE: a collaborative engine for practical disease prediction. Data Mining and Knowledge Discovery 20 388-415.