Implementation of a Recommender System on Medical Recognition and Treatment

Meisamshabanpoor and Mehregan Mahdavi

***Abstract*—On the recent decade, penetrating of computer systems on medical decisions has beenone of the most important issues in medical science. This paper qualifies a medical recommender system for disease recognition and treatment. So, wewill have a look to recommendation concept and its definition.After a total view, the steps of implemented recommendation system and its algorithm will be represented.**

***Index Terms*—Recommender systems, symptoms, characteristics, partial group**

I. Introduction

One of the software systems that have been enhancing in participation on recommendations and making decisions is recommender system.These systems are able to adopt, select or recommend one element between big volumes of relative information [1]. Thispaper talks about a recommendation method to recognize and treat diseases by classification. This system is able learn by the information cashed from patients.

II. Recommender System Foundation

In the age of information overload, people use variety of methods and strategies to make choice between many items. Thischoice can be about what to buy, how to spend leisure time, which item is consistent with them. What should they use to solve their problem? Recommender systems made some of these subjects and case studies automatic and optimized. [2]

Collaborative Filtering

Collaborative filtering is a method that we will use to filter.Collaborative filtering is a technique that makes a filtered set of the data. It helps us to have an abstract view of relative data.Today, systems of this kind are in wide use and have also been extensively studied over the past fifteen years. [3]

III. Master Frame Work and the Target of The Project

This project is a medical recommendation system to disease recognition and appropriate medication by specifying its requirements and functional structure. The learning on medical system has been always an important feature that has a wide scope of research. Hence, we tried to make a relative learning step on this system. The last feature of this system is its anticipation capability for the required period of treatment.Based on this matter, the system should anticipate how long is needed to finish the treatment.

IV. Total View

Foundation of this system is situated on a List.This List has two dependentparts which contains different type of information. One partcontainsessential characteristics and another one is digit information scope. These two sets of information are situated on two different levels. Indeed, digit information is located in a level after characteristics. Each of them has its own effects and functionalities.

V. Characteristics

Physicians use different characteristics to make some specific categories of patient to analyse and recognize diseases. On this project we use these elements: Patient ID(PID), Age, *BMI* .We use Age and *BMI* on two levels: Firstly, we use Age to make a separated category based on the age of patients. Here, every 10 years make a group. Secondly, BMI should be used as an important item to make better classification with further details. *BMI* is the standard offitness and tell that: $BMI= H - (W*W)$

On this equation H is the patient high (Meter) and W is patient weight. *BMI* makes us independence from relation between age and weight.BMI measurement is:[4]BMI<20=Thin, 20<BMI<25=Suitable weight, *25<BMI<30= Extra weight*, 30<*BMI*= *Fat.* Interaction between these two levels makes partial groups. However, we can append or delete some of the elements to gain stunning sensitivity on the algorithm. Now, we have a classified environment which situates every person in a particular category. (Fig. 1)

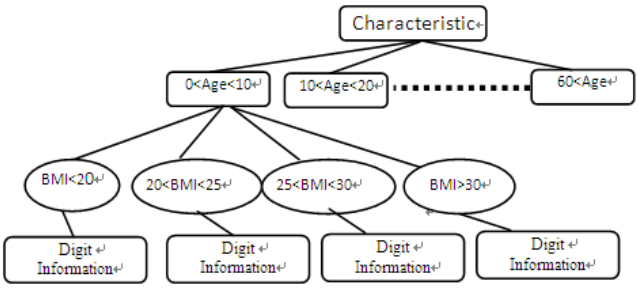

Fig. 1. Classified elements of characteristics.

Manuscript received May 15, 2012; revised July 6, 2012.

Meisamshabanpoor is with the Department of Information Technology International Campus University of Guilan, Iran (e-mail: shabanpoor@gmail.com).

Mehregan Mahdavi is with the Department of Computer Science and Engineering University of Guilan, Iran.

DOI: 10.7763/IJEEEE.2012.V2.136

VI. Digit Information

Digit information is arithmetic part of List which has direct impression on the recognition and treatment. It should be implemented on each partial group. It has three attributes:

Symptoms of Disease, array s, Name of Disease, Medicines

We should cast our plan and implement mathematical structure in this part of the List. The list is a set of digit elements between *1...n*. All of the possible symptoms for specific area of human's body should be considered on this part. Numeric value of each symptom is a number confined between 0.5. This means that if *s[i] =5*, [5] symptom i is absolutely appeared for a specific patient, if *s[j] =0*, the symptom j is not appeared on the patient.

VII. System Implementation

Now we have some basic variables. The k is the number of possible diseases can be appeared on specific parts of patient's body and m is explicit and initial number of default patients and *m=k*. We explain them as fallowing: *S= [$s_1 .. s_n$]* of symptoms, *P= [$p_1 .. p_m$]* of patient, *D= [$d_1 .. d_k$]* of diseases. Since on the first time of use we do not have any main patient, so number of system's cells can be calculated as following: $Cells = 35 \times m \times (n + 3)$

The 35 is set of number of partial groups gained by number of age groups × number of *BMI* classes and 3 is the columns *PID*, Name of Disease and Medicines. Totally for ever Pi, Sjwe will have:

r= amatrix of ratings ri,j for each partial group withi:*1 . . . m, j:1 . . . n*

VIII. Filtering

As a case study, assume a patient as main patient attends to use system. Recognition what partial group is related to the main patient first step can be easily found out. After partial group exploration, employing some techniques for comparison is main subject. To gain the result, main patient's related information should be compared with default patient. On the other side, some people tend to give only high ratings, whereas others will never give a 5 to any symptom. On the other side, there is some people use small number to rate. So it can be a critical section of the algorithm. A referable approach should integrate all of this information and also difference of rating methods should not impress the results of the system. One common measure used in recommender systems is Pearson's correlation coefficient. The similarity *sim(a, b)* of patients *a* and *b*, given from the rating matrix *R*, is defined in formula. The symbol $\bar{r}_a$corresponds to the average rating of patientato symptoms:

$$\bar{r}_a = \frac{\sum_{s \in S} r_{a,s}}{n}$$

The Pearson coefficient factors these averages out in the calculation to make patients' situation comparable. In fact it ignores many noise of the system caused by variety in patient ratings. [6]

$$Sim(a,b) = \frac{\sum_{s \in S}(r_{a,s} - \bar{r}_a)(r_{b,s} - \bar{r}_b)}{\sqrt{\sum_{s \in S}(r_{a,s} - \bar{r}_a)^2} \cdot \sqrt{\sum_{s \in S}(r_{b,s} - \bar{r}_b)^2}}$$

The Pearson correlation coefficient takes values from +1 (strong positive correlation) to−1 (strong negative correlation) corresponde to data points lying exactly on a line (in the case of the sample correlation), or to a bivariate distribution entirely supported on a line (in the case of the population correlation). Thus, nearing toward + 1 can be criteria to choose a neighbour. [7] In some application of recommender system that users are free to fill the values of matrix. This coefficient does not have enough productivity; because there are many items do not have rating values and this influences the results of the formula and makes a frail accuracy. But on this system, all symptoms should be filled and at least gain 1 as a value, so, there is nothing to be said about disability of the correlation. Hence, we do not face sparse matrix and can catch stunning accuracy. We use a small partial group (table1):

TABLE I: Ratings Database for Collaborative Recommendation

PID	Symptoms of disease				Disease name
	S[1]	S[1]	S[1]	S[n]	
P1	3	1	2	3	D1
P2	4	3	4	5	D2
P3	3	3	1	3	D3
P4	1	5	5	3	D4

Now David as main patient starts to use the system. He has these rates: *S[1]=5* , *S[2]=3* , *S[3]=4* , *S[4]=4.* Firstly to clarify similarity, we use Pearson correlation to make a unified set of number between -1 and +1.

$\bar{r}_a = \frac{\sum_{s \in S} r_{a,s}}{n}$, $\bar{r}_{david} = \bar{r}_a = 4$, $\bar{r}_{p1} = \bar{r}_b =$ *2),* $Sim(a,b) = \frac{(5-\bar{r}_a)*(3-\bar{r}_b)+(3-\bar{r}_a)*(3-\bar{r}_b)+\cdots+(4-\bar{r}_a)*(3-\bar{r}_b)}{\sqrt{(5-\bar{r}_a)^2+(3-\bar{r}_a)^2+\cdots}\sqrt{(3-\bar{r}_b)^2+(1-\bar{r}_b)^2+\cdots}} \approx$ 0.82

The similarities to the other patients, *p2* to *p4*, are 0.5, 0.00, and − 0.73, respectively. Based on these calculations, we observe that *p1* and *p2* were somehow similar to David in their rating behavior, which both illustrates the similarity between David and *p1* and the differences in the ratings of David and *p4.*

IX. Neighborhood

Now we should choose neighbours. Important thing on selecting neighbourhood is the size of neighbours, because taking all neighbors into account increases the complexity and response time. Now we can define a specific minimum threshold of user similarity or to limit the size to a fixed number and to take only the k nearest neighbors into account. The potential problems of either technique are discussed by Anand and Mobasher (2005) and by Herlocker et al. (1999) they expressed to important facts: [8]

If the similarity threshold is too high, the size of the neighbourhood will be small for many patients, which in turn means that for many items no predictions can be made (reduced coverage).

In contrast, when the threshold is too low, the neighborhood sizes are not significantly reduced.

The value chosen for k – the size of the neighborhood – does not influence coverage. When the number of neighbors k taken into account is too high, too many neighbors with limited similarity bring additional "noise" into the predictions. When k is too small – for example, below 10 in the experiments from Herlockeret al. (1999) – the quality of the predictions may be negatively affected. An analysis of the MovieLens dataset indicates that "in most real-world situations, a neighborhood of 20 to 50 neighbors seems reasonable" (Herlocker et al.2002).

So, we use a threshold on the Pearson Coefficient. We use criterion that if any of Pearson criterion is not positive; alter it by -1. Then eight boursare taken into account if their Pearson Coefficient is not -1.

X. Prediction

On this step we have to decide which of the neighbor's ratings we shall take into account and how strongly we shall value their opinions. On the different types of subject, this step is time consuming with complexity. But on this subject, mentioned problem does not exist and it has just a simple comparison process. This comparison will be executed between every selected neighbor and main patient. Apparently, one of those is interested who has biggest Pearson correlation Coefficient. [8]

XI. Learning Step

We expect that system learn step by step and optimize itself. So, the final learned classes are situated at one step after partial groups and they are the subsets of partial group. There are two different attitudes for each final related partial group: on the first one, all symptoms of main patient and default patient are same. So there is noneed to update the list. Secondly, There are some difference symptoms between two patients-default and main. So, according to their similarity, a class should be made by them. This class shows groups of patients that have same disease with a little difference.

XII. Period Prediction

Next level of calculation is about the time of period needed for patient recovering and treatment. As a simple fact, the rate of a symptom has direct impression on the required time of treatment. For the first patient we do not have an accurate measure for required time to recover. So, we use the upcoming patients to learn required time. On the other word, after finding stunning class for patient we use this steps and process:

Allocate medicines to the first patient

Use the time spent to recover for that patient as default time W(times of week).

For each symptom of the patient use this formula:

$$C_{a,s_i} = \frac{r_{a,s_i}}{\sum_{i=1}^{n} r_{a,s_i}} \times 100 \quad (1<i<n)$$

//effect percentage of each symptom on sum of rate

$rd_{a,s_i} = w \times C_{a,s_i}$//symptom effect on the recovering period

On the next step, we must estimate patient brating's effect for every symptoms. This target can be achieved by using a rating number. ES is estimated effects of every symptom for b. The ES values must be calculated for all symptoms. The every obtained value is a part of w of patient a and is not adopted with *b*. So, we can you use $(\mathrm{Sim}(\mathrm{a_i}, \mathrm{b_i}))$to make consistent values for *b*.

$$ES_{b_{s_i}} = {(C_{a,s_i} \times C_{b,s_i})}/{rd_{a,s_i}},$$

Required time=$\sum_{i=1}^{n} ES_{b_{s_i}} \times sim(a,b)$

As an instance, assume treatment duration of David with the above mentioned inductive data is 5 week=35 days. Now we have a new patient Peter that is situated on the same class with David with this rating number:

S[1]=4 , S[2]=2 , S[3]=3 , S[4]=2 , $\bar{r}_{david} = \bar{r}_a = 4$, $\bar{r}_{peter}$=$\bar{r}_b$= 2.7), $Sim(a,b)$ =0.85

C_{a,s_1}=0.32, C_{a,s_2}=0.18, C_{a,s_3}=0.25, C_{a,s_4}=0.25

rd_{a,s_1} =1.55 W, rd_{a,s_2} =0.9 W, rd_{a,s_3} =1.25 W, rd_{a,s_4}=1.25 W

C_{b,s_1}=0.37, C_{b,s_2}=0.18, C_{b,s_3}=0.27, C_{b,s_4}=0.18

$ES_{b_{s_1}}$ = 1.85 week ≈ 13 days, $ES_{b_{s_2}}$ = 0.9 week ≈ 6 days,$ES_{b_{s_3}}$ = 1.35 week ≈9 days, $ES_{b_{s_2}}$ = 0.9 week ≈ 6 days,

Required time=$\sum_{i=1}^{n} ES_{b_{s_i}} sim(a,b)$=1.85×0.85+0.9×0.85+1.35×0.85+0.9×0.85≈4.24 week≈30 days

XIII. Total algorithm

According to mentioned process we can represent a total algorithm as a helpful total view for implementing.

```
Defining partial groups
List record of arrays contains:
            Characteristic record of:
                  Age: array [1..7] from BMI
      BMI: array [1..5]from BMI classes
            Digit information record of:
   S= [s1..sn] of symptoms
   P= [p1..pm] of patient
   D= [d1..dk] of diseases
Threshold

            m=k

For each partial group
//data entry for m member of Digit information's attributes side of the list
//Input information ofa as main patient
//Pearson correlation coefficient

            $\bar{r}_a = \frac{\sum_{j=1}^{n} a_j}{n}$

For i=1 to m
      {   $\bar{r}_{b_i} = \frac{\sum_{j=1}^{n} p_{i,j}}{n}$

            For j=1 to n
{$Sim(a,b_i) = \frac{\sum_{s\in S}(r_{a,s}-\bar{r}_a)(r_{b,s}-\bar{r}_{b_i})}{\sqrt{\sum_{s\in S}(r_{a,s}-\bar{r}_a)^2}\cdot\sqrt{\sum_{s\in S}(r_{b_i,s}-\bar{r}_{b_i})^2}}$
If $Sim(a,b_i) < 0.1$ then $Sim(a,b_i) = -1$ }
For $i$=1 to $t$ // t PID with the positive $Sim(a,b_i)$
{Max=$p$(1) If $P$($i$)> max then max= $p$($i$)}
Returnp($i$)} //make a CALSS with the $P$($i$) and main patient
If class_member>2 then    //class_member is a function returns the number of the class's members
{Max_ sim    //Max _sim is a function returns the PID that has maximum value called a of similarity with $b$
For $i$=1 to $n$
    $C_{a,s_i} = \frac{r_{a,s_i}}{\sum_{i=1}^{n} r_{a,s_i}} \times 100$
    $rd_{a,s_i} = \mathrm{w} \times C_{a,s_i}$
    $ES_{b_{s_i}} = {(C_{a,s_i} \times C_{b,s_i})}/{rd_{a,s_i}}$
    $RT$=$\sum_{i=1}^{n} ES_{b_{s_i}} \times sim(a,b)$ //requested time}
```

XIV. The Important Challenges

During the eras, the medicine has been always on the promotion line. There are many tools made by fundamental sciences which are helpful. Nowadays, there are more needs. In this paper we tried to make a system that can accomplish as a physician. However we tried to make an applicable recommender system, but there are many challenges which should be attended to make a better system. As a sample, we know that each symptom has its own effect on a specific disease and does not have equal importance with other symptoms. Also, decreasing rate of every symptom is unique. On the other word, all symptoms will not be decreased with same pattern. So, we need to use some coefficient for every symptom depending on its type. This is one bridge between to sciences information technology and medical sciences. [8]

References

[1] G. Adomavicius and Y. O. Kwon, "New recommendation techniques for multicriteria rating systems," *IEEE Intelligent Systems*, vol. 22, no. 3, pp. 48–55, 2007.

[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*, vol. 17, no. 6, pp. 734–749, 2005.

[3] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*, Addison-Wesley, 1999.

[4] F. Belanger, "A conjoint analysis of online consumer satisfaction," *Journal of Electronic Commerce Research*, vol. 6, pp. 95–111, 2005.

[5] S. Berkovsky, Y. Eytani, T. Kuflik, and F. Ricci, "Enhancing privacy and preserving accuracy of a distributed collaborative filtering," Proceedings of the 2007ACMConference on Recommender Systems (RecSys '07) (Minneapolis),ACM, 2007, pp. 9–16.

[6] J. Bettman, M. Luce, and J. Payne, "Constructive consumer choice processes," *Journal of Consumer Research*, vol. 25, no. 3, pp. 187–217, 1998.

[7] D. Boneh and M. K. Franklin, *Efficient generation ofshared RSA keys*, in CRYPTO'97, pp. 425–439, 1997.

[8] P. Bunn and R. Ostrovsky, *Secure two-party k-meansclustering*, in CCS'07, pp. 486–497, 2007.