# SYMPTOM BASED DISEASE PREDICTION AND MEDICINE RECOMMENDATION SYSTEM

Dr.T.Venkat Narayana Rao
Professor , Department of C.S.E
Sreenidhi Institute of Science and Technology, Yamnampet , Hyderabad, India.

Anjum Unnisa, Kotha Sreni & Ruchika Rachakonda
Department of C.S.E
Sreenidhi Institute of Science and Technology, Yamnampet , Hyderabad, India.

## Abstract:

Health and medicine is gaining a lot of importance in today’s advancing world, where evolving technology is being used to combat almost all the known diseases. However, according to reports, more than 200 thousand people in China and 100 thousand in the USA, die each year due to medication errors. Technologies such as data mining and recommender technologies provide possibilities to explore potential knowledge from diagnosis history records and help doctors to diagnose the medical disease and prescribe medicines correctly to decrease medication error effectively. This project proposes a system which takes as input the symptoms of the patient to predict the disease, which is followed by recommending the appropriate medicine. This system consists of a database system module, data preparation module, disease prediction module, medicine recommendation module, model evaluation and data visualization module. A Decision tree map, Naive Bayes model and Random Forest algorithm are used to achieve the objective. This paper deals with the implementation of a system which performs the dual function of prediction of diseases and the recommendation of medicines, which adds to the capabilities of the existing systems.

**Keywords**: Recommendation system, Prediction, Naive bayes, Random forest, Decision tree.

## I. Introduction

It is estimated that more than 70% of people in India are inclined to general body maladies like viral, flu, cough, cold etc, in every 2 months. Since numerous individuals don't understand that the general body illnesses could be side effects to something increasingly hurtful, 25% of the populace surrenders to death because of ignoring the early general body symptoms[1]. This could be a risky circumstance for the population and can be alarming. Hence recognizing or predicting the disease at the earliest is essential to maintain strategic distance from any undesirable losses. The currently available systems are the systems that are either devoted to a particular sickness or are in the research phase for algorithms when it comes to generalized disease.

This situation is not only limited to India, but similar observations are seen across the world, even in countries with robust healthcare systems. According to reports, more than 200 thousand people in China, even 100 thousand in the USA, die each year due to medication errors. In addition to this, various studies show that almost lakhs of people die due to the medication errors[2]. These errors can be attributed to doctors, who prescribe medicines based on their experiences.

Technologies such as data mining and recommender systems give prospects to investigate, provide opportunities to inspect potential knowledge from historical records pertaining to diagnosis and help specialists to analyze the clinical malady and endorse prescription accurately to decrease medication error effectively.

## II.RELATED WORK

The traditional way consists of doctors performing a patient's diagnosis and prescribing medication by virtue of the doctor's experience. This may sometimes lead to the doctors prescribing wrong medicines or an overdose to patients, which causes more health issues to the patients.

Research has been done by many people to build models which can predict diseases. D. A. Davis, N. V. Chawla, N. Blumm, N. Christakis, and A.L. Barabasi published a paper titled "Predicting individual disease risk based on medical history,"[3] which included a novel system named CARE that combined collaborative filtering methods with clustering to predict each patient's greatest disease risks based on their own medical history and those of similar patients.

After the first step of accurate diagnosis has been completed, one can pay attention to the progression of the disease. Understanding how the disease progresses is more important for proactive healthcare.

Another paper "A Bayesian learning approach to promoting diversity in ranking for biomedical information retrieval," by X. Huang and Q. Hu, described the term "medical retrieval" as the dominant way for knowledge exchanging and sharing[4]. Huang etal.proposed a re-ranking model for promoting diversity in medical search.

## II. THE PROPOSED SYSTEM

Through our project, we propose a system which takes as input the symptoms of the patient to predict the disease, which is followed by recommending the appropriate medicine[5]. To simplify the task of the user, instead of having to answer multiple questions which normally constitutes a consultation, the user will simply have to enter the symptoms they are exhibiting. The medical data regarding the symptoms will be stored as a dataset. The task of the system is to respond to the query given by the user by employing a suitable machine learning model on the dataset.

The symptoms of the patients are taken by the doctors for the continuous evaluation of vitals like heart rate, blood pressure, sugar level etc for the analysis[6]. A doctor can search in the system or can fire questionnaires to the system. The system will respond according to the corresponding dataset. This system is mainly designed to help doctors integrate prediction modules and recommendation modules so that it can recommend medicines based on the respective disease, and thereby constitute a thorough system.

The framework employed mainly consists of six modules, as shown in figure 1.

1. Database System Module
2. Data Preparation Module
3. Disease Prediction Module
4. Medicine Recommendation Module
5. Model Evaluation Module
6. User Interface

A Decision Tree map, Naive Bayes model and Random Forest algorithm are used to achieve the objective of disease prediction and medicine recommendation. Since high accuracy and potency is important for such an symptom based disease prediction and medication recommender system[6][10].

, thus we tend to assess some data processing approaches to get an honest trade-off among the accuracy, efficiency and quantifiability.

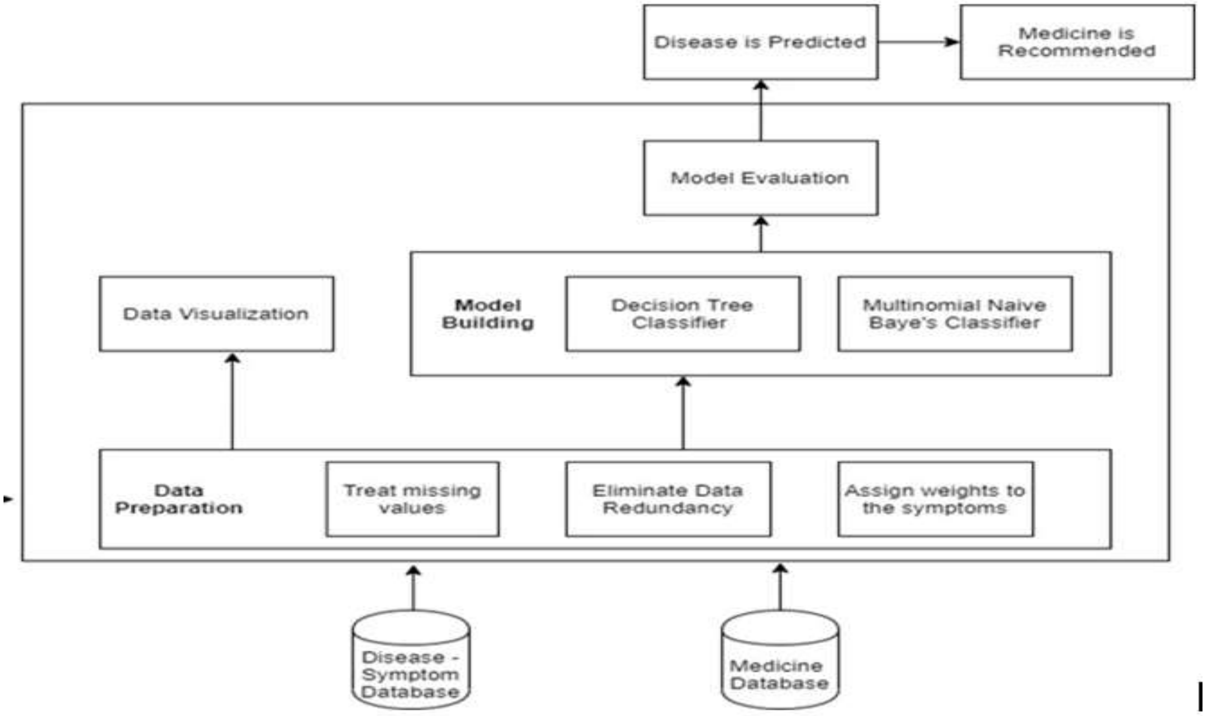


**Fig.1. Architecture of the System**

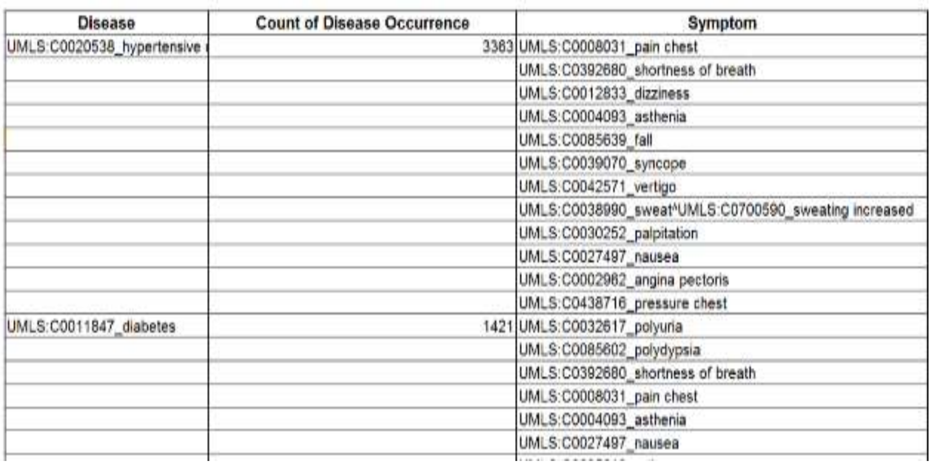
### 1. Database system module:

Two different datasets are used for accomplishing the objective.

The first dataset i.e **Disease-Symptom dataset is a** knowledge database of disease-symptom associations generated by an automated method based on information in textual discharge summaries of patients at New York Presbyterian Hospital admitted during 2004 which includes diseases and their associated symptoms. The objective of this dataset is to predict the illness the user might be suffering from based on the symptoms they exhibit. The attributes are *'Disease Id'*, *'Disease'*, *'Symptoms'*, and *'Count of Disease Occurrence'*. *'Disease Id'* helps in identifying each disease uniquely. The *'Count'* attribute tells us which diseases occur more frequently and thus have a higher chance of occurrence with regard to novel cases.

| **Attribute Name** | **Attribute Type** |
|---|---|
| Disease | Categorical |
| Count of disease | Numerical |
| Symptoms | Categorical |

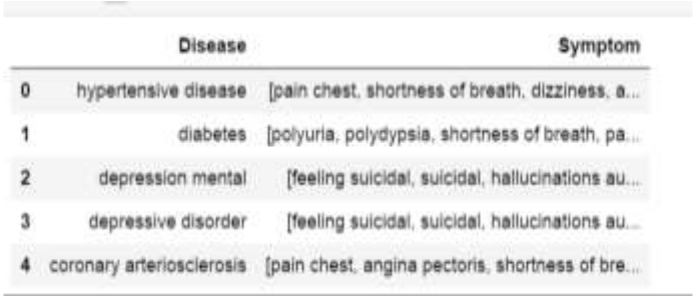
**Table 1:Dataset Description**



| Disease | Count of Disease Occurrence | Symptom |
|---|---|---|
| UMLS:C0020538_hypertensive | 3363 | UMLS:C0008031_pain chest |
| | | UMLS:C0392680_shortness of breath |
| | | UMLS:C0012833_dizziness |
| | | UMLS:C0004093_asthenia |
| | | UMLS:C0085639_fall |
| | | UMLS:C0039070_syncope |
| | | UMLS:C0042571_vertigo |
| | | UMLS:C0038990_sweat^UMLS:C0700590_sweating increased |
| | | UMLS:C0030252_palpitation |
| | | UMLS:C0027497_nausea |
| | | UMLS:C0002962_angina pectoris |
| | | UMLS:C0438716_pressure chest |
| UMLS:C0011847_diabetes | 1421 | UMLS:C0032617_polyuria |
| | | UMLS:C0085602_polydypsia |
| | | UMLS:C0392680_shortness of breath |
| | | UMLS:C0008031_pain chest |
| | | UMLS:C0004093_asthenia |
| | | UMLS:C0027497_nausea |

**Fig 1.1 Raw Dataset**

The above figure 1.1 shows the dataset extracted from the archives of New York Presbyterian Hospital. This data needs to be processed to make it suitable for computation.

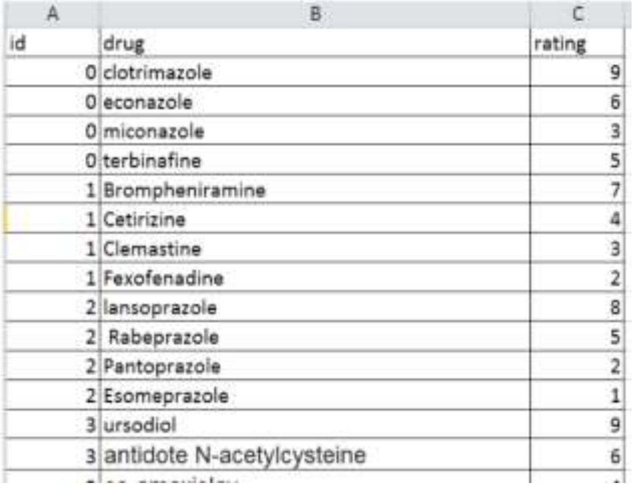


| | Disease | Symptom |
|---|---|---|
| 0 | hypertensive disease | [pain chest, shortness of breath, dizziness, a... |
| 1 | diabetes | [polyuria, polydypsia, shortness of breath, pa... |
| 2 | depression mental | [feeling suicidal, suicidal, hallucinations au... |
| 3 | depressive disorder | [feeling suicidal, suicidal, hallucinations au... |
| 4 | coronary arteriosclerosis | [pain chest, angina pectoris, shortness of bre... |

**Fig 1.2 Disease-Symptom Dataset**

The above figure 1.2 shows the processed dataset is composed of disease names and all the associated symptoms exhibited by a person suffering from that disease.

The second dataset is the **Disease-Drug dataset.** Once the disease has been predicted, i.e, the patient has been diagnosed correctly, the next step is to aid the recovery process by prescribing the right kind of medication. This dataset contains *'Disease Id'*, *'Drug'* and *'Drug Rating*' as attributes. The attribute *'Disease Id'* acts as the link between the two separate processes of disease prediction and drug recommendation. The *'Drug'* attribute enlists the name of the drugs used for treating the corresponding disease. The *'Rating'* attribute is based on the user-opinion and gives information about the goodness or efficiency of the drug.



| A | B | C |
|---|---|---|
| id | drug | rating |
| 0 | clotrimazole | 9 |
| 0 | econazole | 6 |
| 0 | miconazole | 3 |
| 0 | terbinafine | 5 |
| 1 | Brompheniramine | 7 |
| 1 | Cetirizine | 4 |
| 1 | Clemastine | 3 |
| 1 | Fexofenadine | 2 |
| 2 | lansoprazole | 8 |
| 2 | Rabeprazole | 5 |
| 2 | Pantoprazole | 2 |
| 2 | Esomeprazole | 1 |
| 3 | ursodiol | 9 |
| 3 | antidote N-acetylcysteine | 6 |

**Fig 1.3 Disease-Drug dataset**

The above figure 1.3 shows the medicine dataset set which contains attributes like unique id of the disease,drug,rating of the drug.

**Steps:**

1. Importing the Disease-Symptom dataset

2. Import the Disease-Drug dataset

3. Describe datasets

**2. Data preparation module:** The Data Preparation Module's primary purpose is to clean the data. It mainly comprises data exploration and data preprocessing. The real-world data is raw in nature and can have many inconsistencies. It can be incomplete, noisy or dirty. Hence, this module is implemented in order to generate clean data.

**Steps:**

1) Preprocessing the Disease-Symptom dataset

1.1 Fill in the missing values

2) Extract the disease names

3) Process Disease and Symptom Names

3.1 Remove disease id and include only name

3.2 Extract the corresponding symptoms

3.3 Get the Disease Names

3.4 Get the Symptoms Corresponding to Diseases

4) Obtain dummy values of corresponding symptoms of a given disease

5) Convert disease names into disease ID's for faster processing

**3. Disease Prediction Module**

Disease Prediction Module deals with choosing the suitable algorithm to be deployed on the data, and is represented in the form of the model.

**Training with partial dataset:**

1) The cleaned data is generally split into two parts i.e., training data and test data.

2) The first fragment of the dataset is called the training data and it is used for developing the model. The second fragment of the dataset is called the test data, and is used as a reference to test the model.

3) A good practice is to choose one third of the dataset randomly to be a part of the training data, while the remaining data forms a part of the testing data. Models like Decision Tree, Random Forest and Naive Bayes were applied on the training dataset.

This was followed by testing of the model by applying it on the testing dataset.

**Training with full dataset:** Three different classification algorithms are imported from the specific libraries and deployed on the dataset. The entire data is fitted into the model. After this, the model is evaluated on the testing dataset.

**Algorithms used:** The algorithms we have used are Decision Tree, Random Forest, and Naive Bayes. They are used to predict the disease in accordance with the symptoms given by the patient to the doctor. The disease is predicted by pressing the '*Diagnose'* button after choosing the symptoms from the drop down menu. The disease along with accuracy can be displayed. The algorithm with highest accuracy is used to predict the diseases.

### 4. Medicine Recommendation Module

Medicine Recommendation Module is used to recommend the most used medicine for the given disease, once it has been diagnosed by the Disease Prediction Module. After predicting the disease, the medicine for that disease is recommended by the system to the doctors[7]. This module helps the doctors to assess the quality of the medicines available so that the manual errors can be reduced[9].

The drug dataset which consists of attributes *Disease Id, Drug,* and *Rating* is used.The task of the project is to find the most suitable drug and then recommend it. The recommendation is done on the basis of the rating of the drug. All the drugs of the particular disease are arranged in an orderly manner based on the rating of the drug. After this, only the drug with the highest rating is included in a newly formed dataset and the rest of the drugs are ignored. This highest rated drug is then displayed to the user when they click the *'Recommend Medicine'* button on the user interface.

**Steps:**

1) Importing the Disease-Medicine dataset

2) Describe the dataset

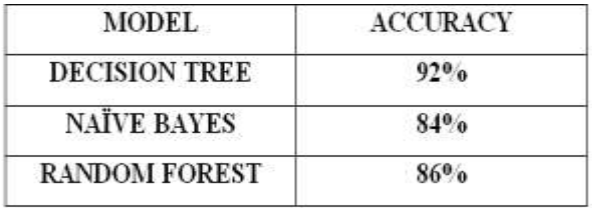
3) Find the highest rated drug for each disease

4) Use the Disease Prediction Module to find out the predicted disease and then map it to the dataset to get the corresponding medicine.

### 5. Model Evaluation

Model Evaluation Module tests the performance of the different prediction and recommendation models that are used on our dataset. The test data is used as a tool to test the accuracy and other related metrics of evaluation. This step determines the precision in the choice of the algorithm based on the outcome[8].

After model building using Naive Bayes, RandomForest, the accuracies obtained is 84%, 86% respectively which are quite low. So in an attempt to improve the accuracy, the project uses Decision Tree as the primary algorithm with 92% accuracy.



| MODEL | ACCURACY |
|---|---|
| DECISION TREE | 92% |
| NAÏVE BAYES | 84% |
| RANDOM FOREST | 86% |

**Table 2:Model Comparisons**

1) Compare the accuracy of 3 models

Accuracy = Number of Correct Predictions / Total Number of Predictions Made

2) Select the model with highest accuracy i.e., Decision tree

3) Display the results of the decision tree model to the user.

**6. User Interface**

**Steps:**

1) Set up a Tkinter window

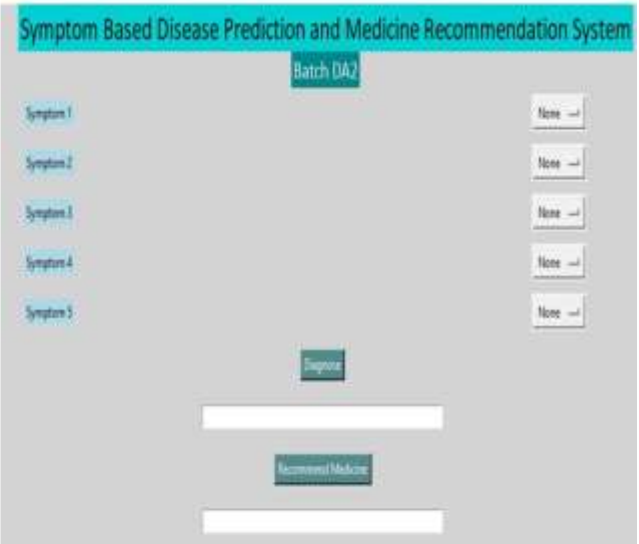
2) Place widgets such as buttons, text-boxes and option menu at their designated positions

3) Give proper functionality to the buttons based on which methods are to be invoked when those buttons are clicked

4) Display the output in a user-friendly manner

## IV. RESULTS AND DISCUSSION

A basic user interface is built for the system which acts as a supporting tool for the doctors and other healthcare workers for the purpose of diagnosis and subsequent prescription of medicines.



Symptom Based Disease Prediction and Medicine Recommendation System

Batch DA2

Symptom 1 None

Symptom 2 None

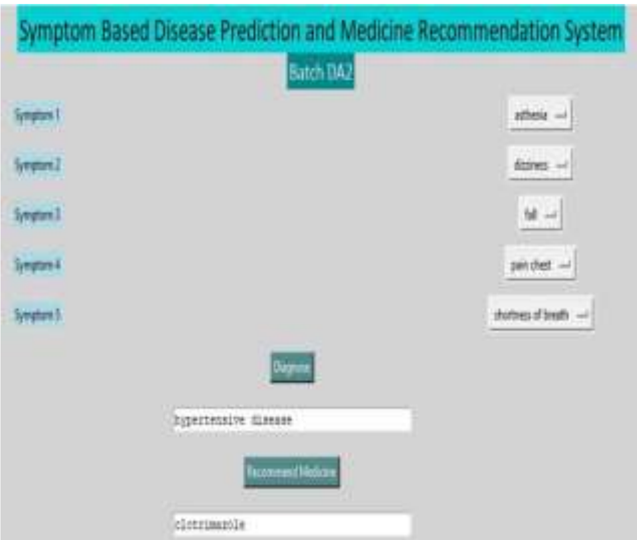
Symptom 3 None

Symptom 4 None

Symptom 5 None

**Fig.5.User Interface**

The above figure shows the user interface of our project i.e “Symptom based disease prediction and Medicine Recommendation System”. Through this interface we can give five symptoms.



Symptom Based Disease Prediction and Medicine Recommendation System

Batch DA2

Symptom 1 asthenia

Symptom 2 dizziness

Symptom 3 fall

Symptom 4 pain chest

Symptom 5 shortness of breath

hypertensive disease

clotrimazole

**Fig.6.Symptom based Disease Prediction and Medicine Recommendation System**

The above figure shows the medicine that can be used to treat the disease which was predicted based on the given symptoms.

## V. CONCLUSION AND FUTURE WORK

We have proposed an idea for a symptom based disease prediction and medicine recommendation system by applying various algorithms on the symptoms and the medicine datasets.This approach is based on four main steps: (i) analysis of symptom and drug dataset (ii) data preprocessing (iii) model building iv)disease prediction and v) Recommending the proper medicine for a particular disease. The proposed system works as a tool for supporting the doctors in their disease diagnosis with the accuracy of 92%. The reliability of the recommendation system may be improved as future work by providing age of the individual, demographic details during the training process. The prescribed medication can also be improved by the brand and the chemical content present in the medication.

## REFERENCES

[1].Manpreet Singh, Levi Monteiro Martins, Patrick Joanis, and Vijay K.Mago,2016, “Building a Cardiovascular Disease Predictive Model using Structural Equation Model & Fuzzy Cognitive Map”,IEEE International Conference on Fuzzy Systems (FUZZ),pp. 1377-1382.

[2].Ashish Chhabbi, LakhanAhuja, SahilAhir, and Y. K. Sharma,19 March 2016,“Heart Disease Prediction Using Data Mining Techniques”, International Journal of Research in Advent Technology,E-ISSN:2321-9637,Special Issue National Conference “NCPC-2016”, pp. 104-106.

[3].Liu Y, Teverovskiy L, Lopez O, Aizenstein H, Meltzer C, Becker J (2007) Discovery of biomarkers for alzheimer’s disease prediction from structural mr images. In: 2007 IEEE international symposium on biomedical imaging, April.

[4] Sathyabama Balasubramanian, Balaji Subramani. SYMPTOM’S BASED DISEASES PREDICTION IN MEDICAL SYSTEM BY USING K-MEANS ALGORITHM, International Journal of Advances in Computer Science and Technology, Vol.3, No.2, February 2014.

[5]. Rahul Isola, Rebeck Carvalho and Amiya Kumar Tripathy.Knowledge discovery in Medical system by using Differential Diagnosis, AMSTAR and K-NN, IEEE Transaction on Information Technology in Biomedicine, Vol.16, No.6, November 2012.

[6]. S.Fox and M. Duggan. Health online 2013.Pew Internet and American Life Project. http://pewinternet.org/Reports/2013/Health-online.aspx,2013.

[7].X. Guo, J. Lu, Intelligent e-government services with personalized recommendation techniques, International Journal of Intelligence Systems, 2007,401–417

[8].T. Lee, J. Chun, J. Shim, S.-G. Lee, An ontology-based product recommender system for B2B marketplaces, International Journal of Electronic Commerce, 2006,125–155.

[9].J.B. Schafer, J. Konstan, J. Riedl, E-commerce recommendation applications, Applications of Data Mining to Electronic Commerce ,Springer, US 2001, 115–153. [5].O.R. Zaiane, Building a recommender agent for e-learning systems, Proceedings of 2002 International Conference on Computers in Education, 2002, 55–59

[10].T.Hung-Wen,S.VonWun,A personalized restaurant recommender agent for mobile e-service, 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service. EEE, 2004, 259–262.