Survey Paper on Recommendation System using Data Mining Techniques

Srinivasa G, Archana M, Patil S S

Abstract— **The purpose of recommendation systems (also known as collaborative filtering systems) is to recommend items which a customer is likely to order. In this paper we describe the recommendation system related research and then Introduces various techniques and approaches used by the recommender system User-based approach, Item based approach, Hybrid recommendation approaches and related research in the recommender system. Generally, recommender systems are used online to suggest items that users find interesting, thereby, benefiting both the user and merchant. Recommender systems benefit the user by making him suggestions on items that he is likely to purchase and the business by increase of sales. we also explained the challenges, issues in data mining and how to build a recommendation system to improve performance accuracy by applying the techniques**

Index Terms— **Recommender system, User Based Approach**

I. INTRODUCTION

Data mining refers to extracting or mining the knowledge from large amount of data. The term data mining is appropriately named as 'Knowledge mining from data' or "Knowledge mining". Data collection and storage technology has made it possible for organizations to accumulate huge amounts of data at lower cost. Exploiting this stored data, in order to extract useful and actionable information, is the overall goal of the generic activity termed as data mining. The following definition is given: Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules. In [1], the following definition is given:

Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful Patterns and rules. Data mining is an interdisciplinary subfield of computer science which involves computational process of large data sets' patterns discovery. The goal of this advanced Analysis process is to extract information from a data set and transform it into an understandable structure for further use. The methods used are at the juncture of artificial intelligence, machine learning, statistics, database systems and business intelligence. Data Mining is about solving problems by analyzing data already present in databases [2].

Data mining is also stated as essential process where intelligent methods are applied in order to extract the data patterns.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

II. LITERATURE SURVEY

In practice, research paper recommender systems do not exist. However, concepts have been published and partly implemented that could be used for their realisation. Some authors suggest using collaborative filtering and ratings. Ratings could be directly obtained by considering citations as ratings [15] or implicitly generated by monitoring readers' actions such as bookmarking or downloading a paper [16], [17]. Citation databases such as Cite Seer apply citation analysis (e.g. bibliographic coupling [18] or co-citation analysis [19], [20]), in order to identify papers that are similar to an input paper [21]. Scholarly search engines such as Google Scholar focus on classic text mining and citation counts. Each concept does have disadvantages, which limits its suitability for generating recommendations. For example [7], citation analysis cannot identify homographs2, and not all research papers are listed in citation databases. Likewise, reference lists can contain irrelevant entries caused by the Matthew Effect [16], self citations [17], citation circles [18] and ceremonial citations6. Recommender systems cannot identify related papers if different terms are used. Collaborative filtering in the domain of research paper recommendation is criticised for various reasons. Some authors claim that collaborative filtering would be ineffective in domains where more items than users exist [22]. Others believe that users would be unwilling to spend time for explicitly rating research papers [15]. Problematic with implicit ratings is that for obtaining the required data, continuous monitoring of the researcher's work is necessary, which raises privacy issues [20]. In general, collaborative filtering has to cope with the possibility of manipulation. Another drawback is that a critical mass of ratings and users is required to receive useful recommendations.

Research survey was conducted to study and classify approximately 96 filtering/recommender systems on various application domains. Out of 96 systems, 21 systems were developed in Web recommendation application domain, 12 systems in movie/TV recommendation application domain, 12 systems in information/document recommendation application domain, eight systems in Usenet news recommendation application domain, seven systems in information filtering and sharing domain, six systems in

Srinivasa G, Asst. prof, Dept of CSE, SVCE, Bengaluru
Archana M, Asst. prof, Dept of CSE, SVCE, Bengaluru
Patil S S, Admin, Dept of CSE, SVCE, Bengaluru

music recommendation domain, four systems in restaurant recommendation application domain, three systems in organizational expertise recommendation domain, three in personalized newspaper domain, three in e-Commerce application domain and software application domain, two systems each in travel recommendation application domain and two in electronic catalogue item recommendation. One system each fall under the recommender application domains such as learning resources recommendation.

III. Challenges and Issues of Recommendation System

3.1 Cold-Start

It's difficult to give recommendations to new users as his profile is almost empty and he hasn't rated any items yet so his taste is unknown to the system. This is called the cold start problem. In some recommender systems this problem is solved with survey when creating a profile. Items can also have a cold-start when they are new in the system and haven't been rated before. Both of these problems can be also solved with hybrid approaches.

3.2 Trust

The voices of people with a short history may not be that relevant as the voices of those who have rich history in their profiles. The issue of trust arises towards evaluations of a certain customer. The problem could be solved by distribution of priorities to the users.

3.3 Scalability

With the growth of numbers of users and items, the sys- tem needs more resources for processing information and forming recommendations. Majority of resources is consumed with the purpose of determining users with similar tastes, and goods with similar descriptions. This problem is also solved by the combination of various types of filters and physical improvement of systems. Parts of numerous computations may also be implemented offline in order to accelerate assurance of recommendations online.

3.4 Sparsity

In online shops that have a huge amount of users and items there are almost always users that have rated just a few items. Using collaborative and other approaches re- commander systems generally create neighborhoods of users using their profiles. If a user has evaluated just few items then it's pretty difficult to determine his taste and he/she could be related to the wrong neighborhood. Sparsity is the problem of lack of information [9].

3.5 Privacy

Privacy has been the most important problem. In order to receive the most accurate and correct recommendation, the system must acquire the most amount of information possible about the user, including demographic data, and data about the location of a particular user. Naturally, the question of reliability, security and confidentiality of the given information arises. Many online shops offer effective protection of privacy of the users by utilizing spe- cialized algorithms and programs.

IV. Methods for Building Recommendation System

The methods used for building recommendation systems rely on machine learning (data mining, statistical inference) techniques. This means that the program is provided with data from the past[1] from which the program should learn to predict the future. Machine learning programs typically work in the following way:

- **Algorithm Design Phase**: One designs a model which can "explain" or fit the data. Usually the designer restricts the model to a specific class of models (for example: decision tree, neural network, probabilistic models etc.) and makes assumptions during the process. Normally, the model has a set of parameters whose values are not specified but are obtained by optimizing a certain cost/loss function. The function that is optimized is application dependent but is usually the training error. In our case the criterion function is the so called root mean squared error (RMSE) which is connected to the standard deviation of the predictions.

Training phase: The program preprocesses the data and estimates the parameters, if any (not all learning methods need to have parameters that are optimized; for example nearest neighbor classifier does not have a training phase).

Tuning phase: In that phase one tests the predictions of the model on the tuning set[2]. If the quality of the results is satisfactory the model is run on the test set and evaluated.

Testing phase: When the model achieves good results from the tuning phase, one can proceed to run the model on the test data. It is important not to use the test data to tune the model's performance. By using the test data too many times one might adjust her model to peculiarities of the test set and obtain results that will not generalize to another test set. This rule is known as "do not train on the test data" and students should not violate it.

V. CONCLUSION AND FUTURE WORK

This paper presented the various techniques to build the recommender system and to improve the performance and accuracy of system. We have also uncovered areas that are open to many further improvements, and where there is still much exciting and relevant research to be done in coming years.

REFERENCES

[1] Warren, K.S. Selective aspects of the biomedicalliterature. *In* Coping with the biomedical literature: A primer for scientists and clinicians, edited by K.S. Warren. 1981. New York, Praeger.

[2] Wurman, R.S. Information Anxiety. 1989. New York, Doubleday.

[3] Mobasher, B.; Cooley, R. & Srivastava, J. Automatic personalization based on Web usage mining. *Communications of ACM*, 2000,

[4] Access log analyzers. Retrieved June 02, 2003 from http://www.uu.se/Software/Analyzers/Accessanalyzers. html

[5] .Gediminas Adomavicius and Alexander Tuzhilin. Using data mining methods to build customer profiles. *IEEE Computer*, 34(2):74-82, February 2001.

[6] Xingquan Zhu, Ian Davidson, "Knowledge Discovery and Data Mining: Challenges and Realities", ISBN 978- 1-59904-252, Hershey, New York, 2007.

[7] Joseph, Zernik, "Data Mining as a Civic Duty – Online Public Prisoners Registration Systems", InternationalJournal on Social Media: Monitoring,