DR B.R. AMBEDKAR NATIONAL INSTITUTE OF TECHNOLOGY JALANDHAR

Assignment – 2

Data Mining and Data Warehousing

SUBMITTED TO-
Dr. Geeta Sikka
CSE Department

SUBMITTED BY-
Ankit Goyal (17103011)
Group - G-1
Branch - CSE

Survey Paper on Recommendation System using Data Mining Techniques

ABSTRACT

The aim of proposed systems (also called collaborative filtering systems) is to suggest items which a client is expected to order. In this paper we describe the recommendation system related research and then introduce 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. Normally, recommended systems are used online to propose items that users discover interesting, thereby, benefiting both the user and merchant Recommender systems benefit the user by building him suggestions on things that he is probable to buy and the business by raising 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.

I. INTRODUCTION

Data mining refers to mining or extracting the knowledge from huge amounts of data. The term data mining is appropriately named as Knowledge mining or Knowledge mining from data. 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 examination and analysis, by semi automatic or automatic means, of huge quantities of data in order to determine meaningful rules and patterns. Data mining is the process of examination and analysis, Semiautomatic or by automatic means, of huge quantities of data in order to determine significant rules and Patterns. Data mining is a subfield of computer science which consists of computational practice of huge data groups‘ patterns finding. The goal of this advanced examination process is to extract data from a data set and convert it into an understandable structure for supplementary use. The methods used are at the moment of artificial intelligence, statistics, machine learning, business intelligence and database systems. Data Mining is concerning solving problems by analyzing data previously present in databases. Data mining is also declared as a crucial procedure where intellectual methods are used in order to extort the data patterns. Data mining consists of five key elements: Mine, transform, and load the operation data onto the data warehouse system.

Save and supervise the data in a multidimensional DBS. Offer data access to information technology professionals and business analysts. Examine the data by application software. shows

the facts in a valuable format, such as a table or graph. This template, modified in MS Word 2007 and saved as a ―Word 97-2003 Document for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons:
(1) ease of use when formatting individual papers,
(2)automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products
(3)conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example.
Some components, such as multileveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

II. LITERATURE SURVEY

In follow, research paper recommender systems do not survive. However, concepts have been available and in part implemented that could be used for their realisation. Some authors suggested using joint filtering and ratings. Ratings could be openly obtained by considering documents as ratings or implicitly generated by monitoring readers‘ actions such as bookmarking or downloading a paper. Citation databases such as Cite Seer apply citation study (e.g. bibliographic coupling or co-citation analysis), in order to recognize papers that are comparable to an input paper. Scholarly search engines such as Google Scholar focal point on typical text mining and citation count Each concept does have disadvantages, which restricts its correctness for generating recommendations. For example , citation analysis cannot identify homographs2, and not all research papers are listed in citation databases. Likewise, reference lists can be full of irrelevant entries caused by the Matthew Effect , self citations [1], citation circles and traditional citations6. Recommender systems cannot identify related papers if different terms are used. Collaborative filtering in the domain of research paper suggestion is criticised for various reasons. Some authors state that collaborative filtering would be ineffective in domains where more items than users exist [2]. Others judge that users would be indisposed to spend time explicitly rating research papers . sticky with implicit ratings is that for obtaining the required data, unbroken monitoring of the researcher‘s work is essential, which raises privacy issues . In general, collaborative filtering has to cope with the opportunity of manipulation. Another negative aspect is that a critical mass of ratings and users is compulsory to receive useful recommendations. Research survey was conducted to study and catalog approximately 96 filtering/recommender systems on various application domains. Out of 96 systems, 21 systems were residential in Web commendation application domain, 12 systems in movie/TV recommendation application domain, 12 systems in information/document recommendation submission domain, eight systems in Usenet news recommendation application domain, seven systems in information filtering and sharing domain,

six systems in music recommendation domain, four systems in restaurant recommendation application domain, three systems in organizational knowledge recommendation domain, three in adapted newspaper domain, three in eCommerce application domain and software application domain, two systems each in travel recommendation application domain and two in electronic catalogue item recommendation. One system each falls under the recommender application domains such as learning resources recommendation.

III. CHALLENGES AND ISSUES OF RECOMMENDATION SYSTEM

A. Cold-Start

It‘s solid to give suggestions to new users as his summarize is almost empty and he hasn‘t valued any items yet so his experience is unidentified to the organization. This is known as the cold start problem. In various systems this problem is solved with analysis when building a profile. Things can also have the cold-start as they are fresh in the system and haven‘t been rated before. Both of these problems can be also evaluated with hybrid approaches.

B. Trust

The attitude expressed of people with a little history may not be that pertinent as the attitude expressed of those who have rich history in their profiles. The issue of confidence occurs towards evaluations of an assured customer. The difficulty can be solved by sharing of facts to the users.

C. Scalability

With the increase of number of items and users, the system requires additional resources for processing data and creating suggestions. many of resources are addicted with the aim of influential users with related goods, and tastes with related metaphors. This problem is solved by the grouping of many types of filters and real progress of systems. Parts of numerous calculations also are implemented offline to speed up assertion of suggestions online.

D. Sparsity

In e-commerce shops that contain an enormous amount of items and users there are almost forever users that contain ratings for a few items. Using shared and other approaches proposed systems generally build neighborhoods of users using their profiles. If a user has examined just little items then it‘s appealing hard to conclude his taste and she/he can be associated with the incorrect neighborhood. Scattering is the difficulty of needing information [3]. E. Privacy Privacy has been the most vital problem. In order to obtain the most correct and accurate suggestion, the system must obtain the mainly quantity of information probable about the user, including data about the location of a specific user and demographic data. Naturally, the question of security, reliability and confidentiality of the given data arises. Various online shops offer efficient security of privacy of the users by using dedicated algorithms and programs.

IV. METHODS FOR BUILDING RECOMMENDATION SYSTEM

The methods worn for building commendation systems depend on machine scholarship (statistical inference ,data mining) practice. This means that the program is given with the data from the previous past in which the program should learn in the route of predicting the future. contraption culture programs typically work in the following way: Algorithm Design Phase: A model which can ―describe‖ or fit the data. regularly the designer restricts the model to a detailed class of models (for example: neural network, decision tree, probabilistic models etc.) and makes supposition during the process. Normally, the model has a set of parameters whose morals are not particular but are obtained by minimizing a certain loss/cost function. The function that is optimized is rationale dependent but is usually the schooling error. In our case the standard function is also known as Root Mean Squared Error (RMSE) which is attached to the standard deviation of the predictions. Training phase: The course preprocesses the data and approximates the parameters, if any (not all learning approaches need to have parameters that are minimized; for example nearest neighbor classifier it does not have training phase). Tuning phase: In that phase one tests the predictions of the model on the tuning set1. If the reputation of the results is acceptable the model is run on the test set and evaluated. Testing phase: When the model achieves good results from the modification phase, one can press forward to run the model on the test data. It is important not to use the test data to tune the model's routine. By using the test data too many times one might regulate their model to peculiarities of the test set and get hold of results that will not oversimplify to another test set. This rule is known as ―do not train on the test data‖ and students should not infringe it.

V. CONCLUSION AND FUTURE WORK

This paper has presented the various techniques to build the recommender system and to advance the performance and accuracy of the system. We have also discovered areas that are open to many additional improvements, and where there is still much exciting and relevant research to be done in future.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, ―On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,‖ Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)
[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, ―Fine particles, thin films and exchange anisotropy,‖ in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.