Volume 2, Issue 5, May 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey of Naïve Bayesian Algorithms for Similarity in Recommendation Systems

Meghna Khatri
Department of Computer Engineering, Maharishi Dayanand University, Rohtak, India
meghna823@gmail.com

***Abstract*— The recommendation systems are widely used to support the users to handle the ever increasing data over the internet efficiently. Recommendation systems apply machine learning and data mining techniques for filtering unseen information and can predict whether a user would like a given resource. To date a number of recommendation system algorithms have been proposed such as collaborative filtering recommendations, content based recommendations and hybrid approach algorithms. The focus is generally on content based recommendation systems methods which are mainly based on naïve Bayesian machine learning algorithm. In this paper, a survey of techniques is presented which suggest naïve Bayesian algorithm for similarity in recommendation systems**

***Keywords*— Collaborative filtering, Content based filtering, Naïve Bayesian algorithm**

I. INTRODUCTION

The data available on the web is of highly dynamic nature especially in the case if e-commerce systems and needs to be handled efficiently in order to provide competitive and efficient applications such as recommendation systems that would be able to predict the ever changing tastes of users accurately.

Collaborative filtering recommender systems recommend items by identifying other users with similar taste and use their opinions for recommendation; whereas content-based recommender systems recommend items based on the content information of the items. Content-based recommendation takes descriptions of the content of the previously evaluated items to learn the relationship between a single user and the description of the new items [2]. In this method a user is supposed to like a new item if the item is similar to other items that are liked by the user [3]. The collaborative recommendations are able to deliver recommendations based on the relevance feedback from other similar users. These systems suffer from scalability, data sparsity, over specialization, and cold-start problems resulting in poor quality recommendations and reduced coverage. Hybrid recommender systems combine individual systems to avoid certain aforementioned limitations of these systems.

Naive Bayesian method is a famous classification algorithm [1] and it could also be used in the recommendation field. When factors affecting the classification results are conditional independent, naive Bayesian method is proved to be the solution with the best performance. When it comes to the recommendation field, naive Bayesian method is able to directly calculate the probability of user's possible interests and no definition of similarity or distance is required, while in other algorithms such as k-NN there are usually many parameters and definitions to be determined manually. It is always fairly difficult to measure whether the definition is suitable or whether the parameter is optimal. Vapnik's principle said that when trying to solve some problem, one should not solve a more difficult problem as an intermediate step. On the other side, although Bayesian network [4] have good performance on this problem, it has a great computational complexity.

New collaborative filtering algorithm are being designed to provide better performance of the algorithm to provide users with more accurate recommendations and consider the basics of naïve Bayesian algorithm for similarity.

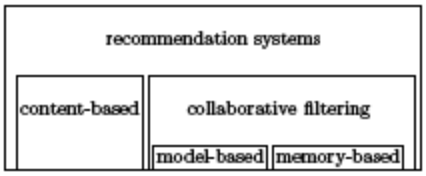
II. BACKGROUND

As shown in Table 1, recommendation systems are implemented in many ways. They attempt to provide items which are likely of interest to the user according to characteristics extracted from the user's profile. Some characteristics are from content of the items, and the corresponding method is called content-based approach. In the same way, some are from the user's social environment which is called collaborative filtering approach [5].

Content-based approach reads the content of each item and the similarity between items is calculated according to characteristics extracted from the content. The advantages of this approach are that the algorithm is able to handle brand new items, and the reason for each recommendation is easy to explain. However, not all kinds of items are able to read.

Content-based systems mainly focus on items containing textual information [6], [7], and [8]. When it comes to movies, the content-based approach does not work. Therefore in this problem, we chose collaborative filtering approach.

Table 1. Various recommendation systems



Compared to content-based approach, collaborative filtering approach does not care what the items are. It focuses on the relationship between users and items. That is, in this method, items in which similar users are interested are considered similar [9], 10]. Collaborative filtering systems try to predict the interest of items for a particular user based on the items of other users' interest. There have been many collaborative systems developed in both academia and industry [9]. Algorithms for collaborative filtering can be grouped into two-general classes, memory-based and model-based [11], [12]. Memory-based algorithms essentially are heuristics that make predictions based on the entire database. Values deciding whether to recommend the item are calculated as an aggregate of the other users' records for the same item. In contrast to memory-based methods, model-based algorithms first built a model according to the database and then made predictions based on the model [12]. The main difference between model-based algorithms and memory based methods is that model-based algorithms do not use heuristic rules. Instead, models learned from the database provide the recommendations.

A Bayesian classifier [34] is a probabilistic framework for solving classification problems. It is based on the definition of conditional probability and the Bayes theorem. The Bayesian school of statistics uses probability to represent uncertainty about the relationships learned from the data. In addition, the concept of *priors* is very important as they represent our expectations or prior knowledge about what the true relationship might be. In particular, the probability of a model given the data (*posterior*) is proportional to the product of the *likelihood* times the *prior probability* (or prior). The likelihood component includes the effect of the data while the prior specifies the belief in the model before the data was observed.Using the Bayes theorem, the probability of a document d being in class Cj is calculated as follows:

$$P(C_j|d) = \frac{P(C_j)P(d|C_j)}{P(d)}$$

where P(Cj |d), P(Cj), P(d|Cj), and P(d) are called the *posterior, prior, likelihood,* and *evidence* respectively.

The Naive assumption is that features are conditionality independent, for instance in a document the occurrence of words (features) do not depend upon each other.

$$P(C_j|d) = \frac{P(C_j)\prod_{i=1}^{h} P(F_i|C_j)}{P(F_1,\cdots,F_h)}.$$

To classify a new document, Naive Bayes calculates posteriors for each class, and assigns the document to that particular class for which the posterior is the greatest.

III. LITERATURE REVIEW

An adaptation of the naive Bayes classification algorithm for the label ranking problem is proposed in [13]. The main idea lies in the use of similarity between the rankings to replace the concept of probability. The proposed method is empirically tested on some metalearning problems that consist of relating characteristics of learning problems to the relative performance of learning algorithms. This method generally performs better than the baseline indicating that it is able to identify some of the underlying patterns in the data.

In other work by [14], a framework for combining the item-based CF with the Naive Bayes classifier. The idea is to use Naive Bayes classifier in off-line stage for generating recommendations. The prediction computed by the item based CF using on-line stage is used if we have less confidence in the prediction computed by the Naive Bayes, else Naive Bayes's prediction is used. We propose a simple approach for determining the confidence in the Naive Bayes's prediction. We compared our algorithm with six different algorithms: user-based CF using Pearson correlation with default voting (UBCFDV) , item-based CF (IBCF) using adjusted-cosine similarity 20 , a hybrid recommendation algorithm, IDemo4, a Naive Bayes classification approach (NB) using item features information, a naive hybrid approach (NH) for generating recommendation 21, and the content-boosted algorithm (CB). Furthermore, we tuned all algorithms for the best mentioning parameters.

A new simple solution to the recommendation topic is provided in [15]. According to our experiment, the improved naive Bayesian method has been proved able to be applied to instances where conditional independence assumption is not obeyed strictly. Our improvement on naive Bayesian method greatly improved the performance of the algorithm. The improved naive Bayesian method has shown its excellent performance especially at long length recommendation.. According to our comparison between ordinary and improved naive Bayesian method, the improvement on naive Bayesian method has an excellent effect. The result of ordinary naive Bayesian method is even worse than that of non personalized recommendation. However, after the improvement the performance is obviously better than the non-personalized recommendation. It is concluded that there is a strong relevance between user's known and unknown interests. The performance of non-personalized recommendation tells that the popular items are also very important to our recommendation. When a proper combination between two aspects is made, as it is in the

improved naive Bayesian method, performance of the algorithm should be satisfactory. When the combination is not proper, it may lead to a terrible performance as it is shown in the ordinary naive Bayesian method.

In another work by [16], an approach is presented that is not limited to any specific recommendation algorithm. The intuition behind this approach comes from the assumption that multicriteria ratings represent user preferences for different components of an item, such as story, acting, direction, and visuals in the case of movies. So, an item's overall rating is not just another rating that is independent of others; rather, it serves as some aggregation function *f* of the item's multicriteria ratings. In other words, this approach assumes that the overall rating has a certain relationship with the multicriteria ratings. For instance, in a movie recommendation application, the story criterion rating might have a very high priority—that is, movies with high story ratings are well liked overall by some users, regardless of other criteria ratings. So, if a system predicts that a movie's story rating will be high, it must also predict that the overall rating will be high in order to be accurate.

In this paper [17], only six hybrid recommendations implementations were shown. However, many other web-based recommender systems were implemented by students of the course "Interactive web-based systems design" and masters works in recent two years, proving that it is possible to implement hybrid recommendation in many different areas. In these implementations the consensus methods were usually used in the collaborative recommendation, but it is also possible to apply the consensus methods for demographic, content-based and also combined hybrid recommendations. It is quite difficult to test recommender systems, especially in controlled condition, because many different users are necessary to show how collaborative method operates, as well as each user needs rather long time of working with the system to show how content-based method operates. However all of the systems were tested with tens of users (one almost 80) and some methods were tested with most exhaustive user tests method, which methodology suggest to test only five users. These tests showed the increase of some usability factors after application of system recommendation.

IV. **CONCLUSION**

As per the requirement of the efficient handling of data over the internet, the improved naïve Bayesian methods provide better performance.

The problems of previous recommendation systems algorithms have been addressed and this approach outperforms others in terms of accuracy, coverage and is more scalable.

The approach used to include naïve Bayesian for similarity in recommendation systems but leave space for suggesting the best performance of these algorithms in terms of time and space complexities.

REFERENCES

[1] Kotsiantis, S.B., Zaharakis, I.D., Pintelas, P.E.: *Machine learning: a review of classification and combining techniques. Artificial Intelligence Review* , 2006

[2] Montaner M,. Lopez B, de la Rosa JL, "*A Taxonomy for Recommender Agents on the Internet*", *Artificial Intelligence Review*, 19, pp. 285-330, 2003,.

[3] Dastani M, Jacobs N, Jonker CM, Treur J, *Modelling User Preferences and Mediating Agents in Electronic Commerce*. LNCS pp. 163-193,1991, 2001,

[4] Yuxia, H., Ling, B.: A Bayesian network and analytic hierarchy process based personalized recommendations for tourist attractions over the Internet. Expert System With Applications 2009

[5] Balabanovic, M., Shoham, Y.: Fab: *Content-Based, Collaborative Recommendation Comm.* ACM 1997

[6] Rocchio, J.J.: Relevance Feedback in Information Retrieval. In: *Salton, G. (ed.) SMART Retrieval System Experiments in Automatic Document Processing*, ch.14. Prentice Hall, Englewood Cliffs 1979

[7] Pazzani, M., Billsus, D.: *Learning and Revising User Profiles: The Identification of Interesting Web Sites*. Machine Learning 27, 313–331 1997

[8] Littlestone, N., Warmuth, M.: *The Weighted Majority Algorithm. Information and Computation* 108(2), 212–261 1994

[9] Adomavicius, G., Tuzhilin, A.: *The next generation of recommender systems: A survey of the state-of-the-art and possible extensions*. IEEE Transactions on Knowledge and Data Engineering 2005

[10] Linden, G., Smith, B., York, J.: *Amazon.com recommendations: Item-to-item collaborative filtering*. IEEE Internet Computing 2003

[11] Breese, J.S., Heckerman, D., Kadie, C.: *Empirical Analysis of Predictive Algorithms for Collaborative Filtering*. In: Proc. 14th Conf. Uncertainty in Artificial Intelligence July 1998

[12] Pernkopf, F.: *Bayesian network classifiers versus selective k-NN classifier*. Pattern Recognition ,January 2005

[13] Aiquzhinov A., Soares C., Serra A., *Proceeding of the 13th international conference on Discovery science* Pages 16-26 Springer-Verlag Berlin, Heidelberg, 2010.

[14] Ghazanfar M. and Pr¨ugel-Bennett A., *Proceedings of the Second International MultiConference f Engineers and Computer Scientists* Vol 1, IMECS March 17-19, ,Hong Kong2010.

[15] Wang K., Tan Y., *Proceedings of the Second international conference on Advances in swarm intelligence* - Volume Part II, Pages 218-227,2011.

[16] Adomavicius G., Kwon Y. *IEEE Intelligent Systems* Volume 22 Issue 3, Pages 48-55 May 2007

[17] Sobecki J.,*International Journal of Computer Science & Applications* ã Technomathematics Research Foundation,*Vol.3,Issue3,pp52-64*,2006

