International Journal of Scientific Research in Computer Science, Engineering and Information Technology

© 2017 IJSRCSEIT | Volume 2 | Issue 4 | ISSN : 2456-3307

A Web Page Recommendation using Naive-Bayes Algorithm in Hybrid Approach

[1]S. Abirami, [2]J. Bhavithra, [3]Dr. A. Saradha

[1,2]Department of Computer Science and Engineering, Dr. Mahalingam College of Engineering and Technology, Pollachi, India

[3]Department of Computer Science and Engineering, Institute of Road and Transport Technology, Erode, Tamunadu, India

ABSTRACT

Web page recommendation has been emerging as a most important application area in mining. In order to predict the users' interests for effective recommendation two methods such as collaborative filtering and content based filtering are considered. Content based filtering is applied by considering information including user's profile and the users' past preferences. User preferences and similarity with other users are considered as primary factor in collaborative filtering method. In probabilistic generative the unobserved user preferences are also considered along with ratings and semantic content. To improve the accuracy and to still improve the user satisfaction this paper applies Naïve- Bayes classifier along with content and collaborative based approach. Naive-Bayes classifier is considered to be more efficient as it considers dynamic and adaptive features for accurate classification. The features that are considered in Naive-Bayes classifier are independent to each other. The performance of the proposed algorithm is measured using the precision and recall.

Keywords : Naive-Bayes Classifier, Content Based Filtering, Collaborative Filtering

I. INTRODUCTION

Web mining is the process where the information is extracted from the web and it can evaluate the effectiveness of particular web site. The information on the web has been increasing, where recommendation should be made effectively. In early days few companies were generating data and others were consuming. Nowadays, all of us were generating data and all of us were consuming. The web mining requires the recommendation system which extracts the required knowledge from the correlated data, since the size of the data is relatively high on the web. [20]

Web recommender system is one which it provides list of web pages that are mostly liked for the users'. The recommender system compares the similar and the dissimilar data among the other content for the effective recommendation. Recommender system gives the list of recommendation using one of content and collaborative based filtering. First, content based approach the recommendation is made using the users' information from their own profile and according to their own interests. For instance, the user likes the web service sa then the services related to that service will be recommended. In Collaborative filtering, the recommendation is made using the interests of other users' having the similar preferences. For instance, if the user likes the service sa and sb. There will be many other user who likes the service sa and sb and also like the service sc. Probably the service sc will be recommended to that user. [19]

The main goal of this paper is to reduce the complexity of handling data and to increase the web service recommendation, which reduces the users' work on giving their preferences (e.g., the user interests and their personal information) and to satisfy their needs. The cold start problem is solved where the user's interests is identified without any information given by user. The accuracy level of prediction for

CSEIT172447 | Received : 12 July 2017 | Accepted : 23 July 2017 | July-August-2017 [(2)4: 141-147]

recommending web page to the users' is increased by applying learning process. The data sparse environment in which the information that are among the user interests are also considered. The sparse data problem for the information in the web is considered where the most related information is recommended nearly and it is handled by identifying the correlation. [3]

In this paper we use Naive-Bayes algorithm for the proposed work to improve the significance of the recommendation system. The main use of this algorithm is to consider many features like keywords, timings, frequency, ratings for the effective recommendation where it is mainly used for classification process. It considers the timings, ratings, keywords and frequency which are taken from the log files. It is highly scalable and it considers number of features as input while processing the data. Example 1.A fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features. [17]

The remaining portion of the paper is organized as follows. Section 2 discusses the related work of this paper. Section 3 explains about the unified collaborative and content based approach. Section 4 introduces our hybrid approach using Naïve-Bayes algorithm. Section 5 reports the experimental results. Section 6 gives the concluding results.

II. RELATED WORK

A. Quality of Service

Quality of service (QoS) is an important factor for describing the web services. The collaborative approach is mainly used for the QoS prediction. The user past experiences are considered from which the nonfunctional characteristics of web services are considered. The work by Zibin Zheng et al. uses collaborative approach in which the missing value prediction is undergoes. The collaborative approach is used from which the web services from the different service users are collected. The QoS value prediction is proposed from which the suitable web services are identified by combining user based and item based collaborative filtering methods. The missing value prediction for the active user is employed for the missing value prediction. From the QoS value prediction the recommendation will become more effective for the active users. [1]

B. Semantic Content

The old method of recommending the user is with historical data, similarity users', based on profile. On the other hand, Semantic content based approach in which the semantic similarity of services is analyzed. The work by Freddy L´ecu´e combines the semantic-content and collaborative based approach in which it considers the semantic similarity of services. The services that are from the past history, similar users and personal information are recommended in which it considers the semantic description of services. The cold start problem is reduced and the sematic accuracy of services are improved. In this semantic based approach the end user past history are collected from which the neighbor users are calculated. The services that are viewed by the neighbor other than the end user viewed is identified. Those services are ranked according to the semantic similarity values and hence the top k services are recommended to the end user. [2]

C. Latent User Preferences

The latent user preferences are considered where the content alone cannot be used to find out the interests about the user hence the rating is also considers where the user unobservable preferences are also considered. Hence the accuracy is increased and the user unobservable preferences are represented. The work by Kazuyoshi Yoshii et al. includes the latent values where the rating and the content data are considered. In order to find the unobserved preferences of users the latent variables are considered. It combines the content and collaborative based approach. The author aim to satisfy the high recommendation accuracy which it recommend the most relevant information and the new-item problem in which it helps the user to find out the appropriate selections that have unfortunately be given few ratings. The latent user preference is used to recommend the users according to their similar interests. The new item problem is that the item which is mostly liked by user but has no rating is solved. The integration of content and rating data will helps to effective recommendation. [6]

D. Cold Start Problem

The cold start problem is one in which the user wants to give more information about them to get the most related information. It can be effectively solved by combining both the content and collaborative based approach. Thus the information can get from the user search history and by considering the similar user the user interests can be identified. The main problem in the recommendation is cold start where it can only be solved by content and collaborative approach. Content based approach is one which the user information is gained from the user profile and their history where the user has to include some information about them personally. Hence the collaborative approach is combined where the cold start problem is solved. [1] [2]

E. Sparse Data Environment

The unified framework considers both the content and collaborative based approach. The document having the latent meaning is also considered. The data which present sparsely that are same as the users, interests is hence recommended. The work by Alexandrin Popescul et al. proposes a generative probabilistic model that incorporates the three-way co-occurrence data among users, items and item content which combines both content and collaborative approach. In three-way aspect model the user who uses the document along with the latent variables, the latent variables are those which are the topics in which the document generates. The user selects according to the topics. It considers the words from the document which the user saw. The k- Nearest Neighbors are used to find the most relevant document which is to be recommended to the new user. The data among the sparse environment is handled and effectively recommended. [3]

III. EXISTING SYSTEM

As web services are increasing it is believed to be one of the standard medium used for data sharing. The web service recommendation is based on two approaches: content and collaborative filtering. The existing system combines both the approaches to overcome some drawbacks in each approach. The probabilistic generative model is proposed in which the rating data and semantic content are considered. The unobservable user preferences are represented by considering the latent variables.

A. Content based approach

The content based approach is used to recommend services to the user. In content based approach the information about the user is collected from their profile. The recommendation is based on their past history, historical data and ratings. By this there will be possibility of user to explicitly provide their information, which leads to the cold start problem. [16]

B. Collaborative approach

The recommendation process in the collaborative approach is made by using the interests of neighbor users. The neighbor user is identified by calculating nearest user with search history similar to active user. The services that are used by the neighbor user are considered. Those web services are listed and ranked according to the active user interests. The new services of the active user interest may be recommended to the user. [16]

C. Probabilistic Generative Model

Probabilistic generative model uses both the rating and the semantic content for the effective recommendation. The latent variables are considered from which the unobservable user preferences are identified. The rating and the content data of the services are considered. The semantic values of the services that are viewed by the user are considered and the latent values that governs the recommendation process. The latent variables are the user preferences. Hence, the unified model of content and collaborative model is effectively used for recommendation process. [16]

IV. SYSTEM ARCHITECTURE

Our hybrid approach includes two main concepts namely, content based approach and collaborative based approach. The mining process mainly consists of three stages: Data collection with preprocessing, pattern discovery and analysis, recommendation. In filtering approach, the pattern discovery is the offline process and the recommendation is the online process. System architecture is depicted in the Fig.1 with recommendation related operations.

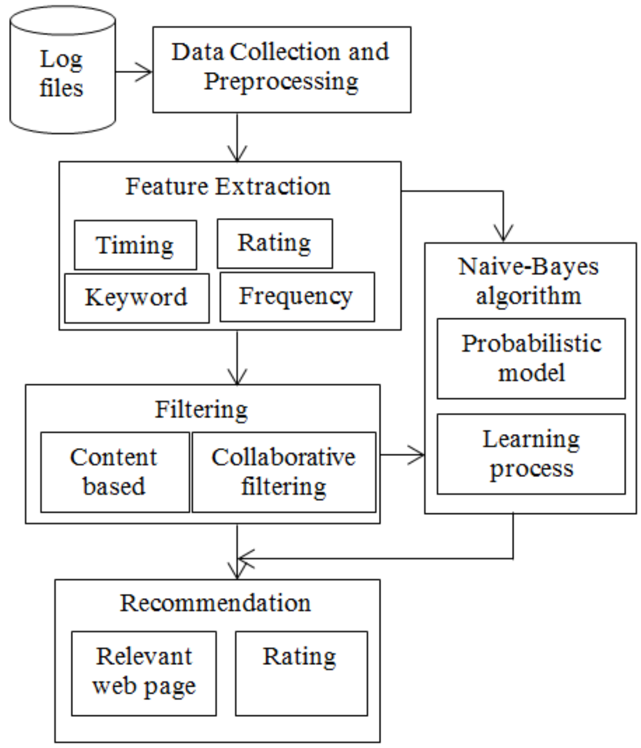

Figure 1. Recommendation using Naïve-Bayes algorithm

The content and collaborative based approach is mainly used for the recommendation process. The main objective is to consider many features like keywords, ratings, frequency for the effective recommendation. The user unobservable preferences are identified.

A. Data Preprocessing & Constructing User Profile

The data preprocessing is a stage in which the information about the users are collected. The log file that has number of users is considered. The log files consist of AnonID, Query, Time and Date, Item Rank, URL about the user.

The dataset is collected from AOL datasets. It consists 45.9 MB of user information. The data set includes {AnonID, Query, Query Time, Item Rank, ClickURL}.

- AnonID - an anonymous user ID number.
- Query - the query issued by the user.
- Query Time - the time at which the query was submitted for search.
- Item Rank - if the user clicked on a search result, the rank of the item on which they clicked is listed.
- Click URL - if the user clicked on a search result, the domain portion of the URL in the clicked result is listed.

The log file is validated hence the unwanted information in the log files are avoided. There are several users and their search information. The users are separated by their ID. Hence, users are separated along with their information using their ID. The separated list consists of information about that user alone.

The user profile constructed is using the content and collaborative based approach. In content based approach the features like keywords, timings, frequency, ratings from each user are calculated.

- Timings: The timing is calculated from the amount of time spent by user for the particular url.

$$\text{Timing}=\frac{\sum_{url \in U}(\text{Closed time url}-\text{clicked time url})}{|U|} \quad (1)$$

- where, U refers to set of url.

- Frequency: The frequency is calculated the amount of time the user visited the url.

$$\text{Frequency}= |U|_{\in u} \quad (2)$$

- where, U refers to number of urls in the url set.
- Ratings: The ratings for the particular web page.
- Keywords: Keywords are those which are extracted from the information in the url. The keywords are extracted by the information in the web page. The url is hence downloaded and the information in the page is preprocessed where the stop words which are the keywords.

In the collaborative approach the neighbor user are identified. The similarity between the users' are calculated in which neighbor users for the active user is identified. The collaborative approach will recommend by comparing with other users. The similarity between the users' is calculated by,

$$\text{Similarity }(u,v)=\frac{\sum_{q \in Q}\sum_{s \in S}\ldots\emptyset(q,s)}{|Q|\ |S|} \quad (3)$$

Where, q and s are queries and services that are viewed by the user from the query set Q and service set S.

Thus the similarity between the users is identified where the neighbor user interests are collected and the services that they liked will be considered other than the same services. Those services are listed and are recommended.

B. Naïve-Bayes Algorithm

Naïve-Bayesian algorithm is an efficient algorithm and which has its independency among each features. It can effectively consider all the features that are extracted from the users' log file which helps to increase the efficiency of the recommendation to the active use. Using Naïve-Bayesian algorithm the url that are to be recommended to the active user will identified by using the features. The similarity measure is calculated between the users' for the better recommendation. The neighbourhood users' are identified by their similarity. The user may visit many web pages to get the accurate information they needed. From those web pages by considering the features the most relevant web pages will be detected and hence they are prioritized. [7] [8] [9]

```
Begin NB Algorithm
Input U: user set
 F: no. of features
 L: no. of url
Output V: url that should be recommended
from the user set the NB is calculated for individual user
for each iteration i є I
 for each url
 for each feature
 compute similarity of a base F1
 if similarity > required similarity add to the expected list.
 For each expected list compute EM
 Sortlist based on the descending order
 Recommend the top k url user based on EM
 Collect feedback and update the similarity
 End for
End NB Algorithm
```

Figure 2. Pseudo code of Naïve-Bayes algorithm.

The pseudo code of the hybrid approach is explained in Fig.2. The content and collaborative based approach is one in which the user interests are obtained. The content based filtering in which the content based filtering the information about the user is collected from individual users' profile and the information provided by them. Collaborative based filtering the user interests is identified by the neighbor users from which the user unobserved preferences are identified. The user having the same interests are grouped as the neighbors where the unobserved preferences are identified.

C. Recommendation Process

The url that are to be recommended will be identified based on ranking and similarity measure. The similarity measure is calculated among the users by comparing their similar interests. The users' are grouped in certain cluster those having the similar preferences. The url that are grouped among the cluster will be sorted in the order. The url is sorted based on the user interests which having the highest priority. The data sparse problem can be removed. Hence, the web page that is to be recommended to the user is sorted where the recommendation will achieve the higher accuracy. The user unobserved preferences are also considered for the better recommendation. The problem of new user in the internet can also be solved using this method. Hence, Naïve-Bayesian algorithm is effective algorithm to recommend the web pages to the user.

V. RESULTS AND DISCUSSION

A. Performance evaluation

The evaluation metrics are used to find the accuracy of recommended website. We use F1 measure as the evaluation metrics. F1 measure is the harmonic mean of precision and recall. [18] F-measure is calculated using:

F1-Measure= $\frac{2}{\left(\frac{1}{P}\right)+\left(\frac{1}{R}\right)}$ (4)

where P and R refers to precision and recall. It is high only when precision and recall is high.

Precision is the ratio of true positive among all retrieved instances. [18] It is calculated using,

Precision= $\frac{TP}{TP+FP}$ (5)

where TP and FP refers to true positive and false positive. TP detects the url when the url is actually present. FP indicates a given url has been interested by the user when it is actually has not been present.

Recall is the ratio of true positive among all positive instances. [18] It is calculated using,

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \quad (6)$$

where FN refers to false negative. FN indicates that a url is not interested while it is actually was interested.

VI. CONCLUSION

In this paper, we use Naive-Bayes algorithm for effective recommendation. It classifies the web site which is to be recommended. To effectively classify the website more feature like keywords, timings, frequency are considered from which the users, interests is identified. In order increase the accuracy the content and collaborative based approach is used. The learning process is used to increase the accuracy of prediction. Hence, our approach is expected that it will accurately predict the interests and recommend it.

Our future work includes that the further recommendation can be increased by using the considering optimization algorithms.

VII. REFERENCES

[1]. Zibin Zheng, Hao Ma, Michael R. Lyu, Fellow, and Irwin King, (2011),"QoS- Aware Web Service Recommendation by Collaborative Filtering",IEEE transactions on services computing, Vol. 4, no. 2, pp. 140–152, June.

[2]. Freddy L´ecu´e,(2010) "Combining Collaborative Filtering and Semantic Content-based Approaches to Recommend WebServices", IEEE Fourth International Conference on Semantic Computing, pp. 200-205.

[3]. Alexandrin Popescul ,Lyle H. Ungar , David M. Pennock,Steve Lawrence, (2001), "Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments", Published in Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, pp. 437444,August.

[4]. Byron Bezerra and Francisco de A. T. E Carvalho, (2004)," A Symbolic Hybrid Approach to Face the New User Problem in Recommender Systems", SpringerVerlag Berlin Heidelberg, pp. 1011–1016.

[5]. Katja Niemann and Martin Wolpers,(2015), "Creating Usage Context-Based Object Similarities to Boost Recommender Systems in Technology Enhanced Learning",IEEE transactions on learning technologies, Vol. 8, no. 3, pp. 274285,September.

[6]. KazuyoshiYoshii,MasatakaGoto,KazunoriKomatani,TetsuyaOgata,HiroshiG.O kuno, (2006),"Hybrid Collaborative and Content-based Music Recommendation Using Probabilistic Model with Latent User Preferences", Published in University of Victoria.

[7]. Meghna Khatri, (2012), "A Survey of Naïve Bayesian Algorithms for Similarity in Recommendation Systems" , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, pp. 217219, May.

[8]. Kebin Wang and Ying Tan, (2011),"A New Collaborative Filtering Recommendation Approach Based on Naïve Bayesian Method", SpringerVerlag Berlin Heidelberg, pp. 218–227.

[9]. Mustansar Ali Ghazanfar and Adam Pru¨gel-Bennett, (2004), "An Improved Switching Hybrid Recommender System Using Naïve Bayes Classifier and Collaborative Filtering",School of electronics and Computer Science,University of Southampton,United Kingdom.

[10]. Mingming jiang,Dandan Song,Lejian liao,Feida Zhu, (2015),"A Bayesian Recommender Model for User Rating and Review Profiling",Tsinghua Science and Technology ,pp 634-643,December.

[11]. Xi Chen, Xudong Liu, Zicheng Huang, and Hailong Sun, (2010), "A Scalable Hybrid Collaborative Filtering Algorithm for Personalized Web Service Recommendation", IEEE International Conference on Web Services,pp-9-16

[12]. Jonathan L. Herlocker and Joseph A. Konstan, Loren G. Terveen, and John T. Riedl, (2004),"Evaluating Collaborative Filtering Recommender Systems", ACM Transactions on Information Systems, Vol. no. 22, pp.5-53,January.

[13]. John Z. Sun, Dhruv Parthasarathy, and Kush R. Varshney, (2014), "Collaborative Kalman Filteringfor Dynamic Matrix Factorization", IEEE Transactions On Signal Processing, Vol.62, pp.3499-3509, July.

[14]. J. Bobadilla, F. Ortega, A. Hernando, A. Gutiérrez (2013),"Recommender systems survey" Published in Elsevier,Universidad

Politécnica de Madrid, Ctra. De Valencia, Spain,pp.109-132,March.

[15]. Mustansar Ali Ghazanfar and Adam Prugel-Bennett,(2010). "A Scalable, Accurate Hybrid Recommender System", IEEE Third International Conference on Knowledge Discovery and Data Mining, pp. 94-98

[16]. Lina Yao, Quan Z. Sheng, Member, IEEE, Anne. H.H. Ngu, Jian Yu, and Aviv Segev(2015), "Unified Collaborative and Content-Based Web Service Recommendation", IEEE Transactions On Services Computing, Vol. 8, No. 3, May/June

[17]. https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[18]. https://en.wikipedia.org/wiki/precision_and_recall_f-measure

[19]. https://en.wikipedia.org/wiki/Recommender_system

[20]. https://en.wikipedia.org/wiki/Web_mining