ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Personalized Context-Aware Mobile Recommendation System Based on Hybrid Filtering Approach.

Sivaramakrishnan N*, Subramaniyaswamy V, Muralidharan K, Dharmarajan V, and Praveenkumar G.

School of Computing, SASTRA University, Thanjavur, Tamil Nadu, India.

ABSTRACT

Over several decades, web search has developed rapidly and mobile internet applications are getting introduced at a faster rate. But due to constraints like bandwidth, it is difficult to retrieve relevant results without parsing through irrelevant information. Furthermore, the current mobile search products are far from being personalized and accurate. Therefore a new combined filtering approach is recommended to remove useless information for personalized mobile search. This can be done by performing hybrid filtering approach. To achieve personalized outcome, we initially predict the user's behaviour using Naive-Bayes classifier. Using Naive Bayes classifier, we develop the user model which filters the users based on behavioural constraint. Filtering is performed for behaviour predicted users only. Firstly, Content based filtering filters the link based on user's previous experience. This filtering uses Term-Frequency Inverse Document Frequency (TF-IDF) to find the most relevant links. These links are passed as input to collaborative filtering. Collaborative filtering approach finds the similar users using clustering algorithms. We have also used Pearson Correlation Coefficient algorithm which finds the extent of similarity between the current user and the similar users. We use User-User Similarity measure to recommend the most popular links suggested by the similar users to the current user. The final outcome is the personalized result for the current User's query.

Keywords: Naive-Bayes Classifier, Content Based Filtering, TF-IDF, Collaborative Filtering, Pearson Correlation Coefficient.

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Given, a particular query searched by multiple users on a standard web search returns same outcome. However, different users expect different outcome depending upon their needs. The user's requirement for personalized information is always present. Generally most of the information displayed are non specific and less relevant. This occurs due to the fact that searching technique employed by most of the standard web search engine do not consider crucial information like user's personal needs and context specific information. A user has to parse through these useless or less useful information to retrieve the information the truly expect from a standard web search.

Thus by going through the user's requirements, it is known that user's context aware information needs are evident. So, the user interface must be able to return context-aware results which the user expects. To achieve this result, we perform the following algorithms: User's behaviour prediction, Clustering and Filtering.

In this paper, we initially predict user's behaviour using Naive Bayes Classifier. This classifier uses probabilistic values of behavioural attributes to predict the user's behaviour. Next we find the user model which is based on specifying constraint on behavioural attributes. Thus we perform filtering for this generated user model. We use Hybrid filtering approach which is a combination of content based and collaborative filtering. The former method uses Term-Frequency Inverse Document Frequency (TF-IDF) to find the most similar links. The outcome of content based filtering is passed to collaborative filtering.

The main objective of collaborative filtering lies in finding the similar users. This is achieved by using Clustering algorithm. K Medoids, one of the clustering algorithms finds the initial cluster center by computing the least cost of clusters for a given K value .This initial cluster center is passed as input to K Means algorithm which finds the clusters (similar users). Additionally, we use Pearson Correlation Coefficient to find the similarity measure for similar users. Once the similar users are found, links are suggested to the current user by using User-User Similarity. The architecture of context aware mobile approach is shown in Fig 1.

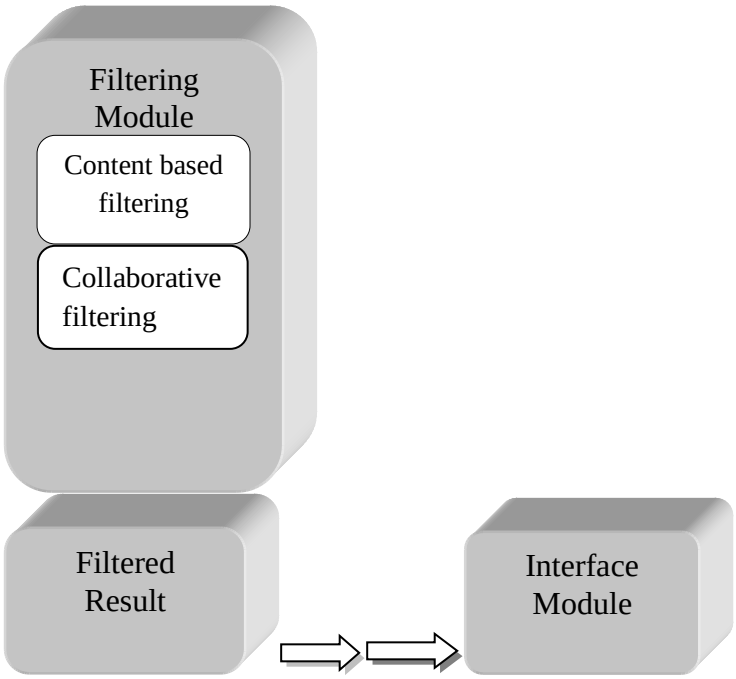


Figure 1: Architecture of context aware mobile search

ISSN: 0975-8585

RELATED WORKS

Naive Bayes Classifier

Krzyszt of Dembczynski et al **[1]** have predicted user behavior based on the real empirical information forms a crucial challenge in data mining. The recorded behaviors of web users are explained using the concept of behavior prediction. The concepts of statistical decision theory along with global models like trend prediction and auto regression are performed to predict the behavior of web users. It is observed that models developed for each user model gives generally better results when compared to global models. Moreover, the time complexity is constant in the number of users and visits and therefore classification does not take significant time.

K. Santra et al **[2]** have proposed Naive-Bayes classifier algorithm for predicting the behavior of interested users instead of making use of upgraded D-Tree algorithm. The concept of Naive-Bayes classification theorem involving probability inferences are used for classification. Results showed that performance metrics i.e. the total amount of time and memory consumed in order to organize the log files using Naïve Bayes classification is very efficient when compared to Decision Tree classification.

Supreet Dhillon et al **[3]** have compared the classification algorithms such as decision tree, Naive Bayes classifier based on accuracy, precision, session based timing and so on. Here the naive part of Naive-Based classifier is assuming word independence that is the chances of particular word in a given category is different from all the probability of words in the same category. This makes Naive Based classifiers more optimal than the effectiveness of other classification algorithms since it does make use of other combinations of words as predictors.

Arne Mauser et al **[4]** predicted the behaviour of the customer by taking the dataset of a German mail order company and splitting it into training and testing datasets. The concept of data pre processing and data classification using Naive bayes and maximum entropy are explained. Maximum Entropy estimates the weights for the posterior distributions for all features in the dataset. We find out that unseen data set are /the best for set of Naive based classifiers with maximum entropy.

Masud Karim et al **[5]** predicted if a person would make a long term deposit by predicting the behaviour of the client using Decision Tree and Naive-Bayes classifiers. UCI Machine Learning dataset is used to investigate the effectiveness of these classifiers. The reason behind this algorithm is the probability of distance in the class of a model. Profit optimal Decision Tree concept is used to get the necessary knowledge. This algorithm obtains actions that change the value of clients from one state to another.

Filtering Techniques

Simon Philip et al **[6]** have integrated the concept of recommendation in digital libraries to trim the amount of incoming information. The similarity value between a paper and a user's profile interest is found using the concept of TF-IFD and Cosine Similarity. The usage of Content-based Filtering technique in recommender systems helps certain users in retrieving those required papers. The result shows that by combining these features in digital libraries will be very helpful to these users. Content-based methodologies are more dependent on that content and less dependent on ratings.

Joonseok Lee et al **[7]** have done a study of comparing collaborative filtering techniques in a variety of experimental contexts and determining its sparsity level, performance criteria and because of that result, we can identify which algorithm works well in what conditions. For Experiments, PREA toolkit is implemented, which implements fifteen Algorithms which includes Constant , User Average , Item Average , User-based , Item-based, Regularized SVD, Bayesian PMF, Slope-One, NPCA, Rank-based CF Algorithms. Various Differences in behaviour has been find out based on the Matrix Factorization methods, which identifies that specific algorithm will do wonder in that particular situation. Matrix – Factorization – based methods have the highest accuracy and all algorithms vary in their accuracy, based on the user count, item count, and density.

Michael J.Pazzani et al **[8]** have recommended the required and relevant web pages from the warehouse. The implementations are done by Content Based Filtering, Collaborative Based Filtering, Demographic filtering. Findings talk about how each Filtering applied in each situation and the drawbacks of individual filtering mechanisms are overcome by the integrated implementation. The Hybrid Approach has given the Best Result over Individual Filtering approaches.

Torres et al **[9]** have applied the Content and Collaborative approaches to get the most required and relevant Research Paper. By Applying hybrid Filtering Mechanisms collectively, the result will be much more efficient. The Relevant Research Papers can be found out with more Efficiency by using the Collective Filtering Mechanisms.

Poonam B. Thorat et al **[10]** have described an overview on recommendation system which suggests the user according to their needs. Filtering is used to provide efficient recommendations to the user. They proposed various techniques of filtering. Firstly, collaborative filtering approach is used which recommends items to the current user, suggested by the similar users. Content based filtering recommends items based on user's previous choice. Finally they proposed a filtering mechanism called Hybrid recommendation which is a combination of above two filtering mechanism. This approach overcomes the problems such as cold start, sparsity problems and it also improves the efficiency of recommendation process.

Clustering

McCallum et al **[11]** have performed clustering on large dataset which are thought to be impossible. The main idea lies in using the distance formula to split the large data into its subsets, which are known as canopies. By implementing canopy approach, computational time on implementing clustering for large data set is reduced.

Sankar Rajagopal et al **[12]** have identified high value and low risk customers by a technique known as customer clustering by using demographic clustering technique. The data is preprocessed and patterns are developed on the data using an algorithm called IBM Intelligent Miner. This is followed by profiling the data, developing the clusters and identifying the low risk and high value customers.

Amandeep Kaur Mann et al **[13]** have described the data mining process which mainly extract useful information from the large data set and make it into understandable form for future use. Clustering is the most important for analyzing and mining the large data. Clustering is done using various algorithms. One of the algorithms is Partition clustering algorithm which produces clusters based on centroid value. Density based clusters defines the area of higher density. The grid based clusters uses single uniform grid mesh to partition the problem domain into cells.

Jadhav Bhushan G et al **[14]** have developed a new search engine using fastest reading algorithm which provide best result. Initially it worked on specific keyword based on text mining. Then base keyword of the content from the database is searched. But the proposed system uses search engine based on text mining and k means clustering. To find keywords using well defined patterns, the K-Means algorithm finds relevant text and implements semi supervised learning clusters algorithm within the database. Thus, these algorithms allow developing an ideal web engine by utilizing the knowledge of database to work with ontology and filtering.

Rakesh Chandra Balabantaray et al **[15]** have performed clustering algorithm to obtain relevant information in the cluster. Once the best clusters are formed, document summarization is executed to focus on key points in a document. They have used hundred of documents to perform clustering algorithms. As a result of clustering documents from various domain can be combined into groups of similar document. Finally, Comparison of k-Means and K Medoids is carried out to find the efficiency of the clusters that are formed.

ISSN: 0975-8585

PROPOSED METHODOLOGY

Naive Bayes Classifier

This method is mainly used for predicting the behaviour based on behavioural attributes of the users. The dataset for this algorithm is split into 2 parts. Training dataset and testing dataset. In the training dataset, the outcome is also given as input to the algorithm. The algorithm calculates probabilistic values for all attributes and predicts the outcome for all users in the testing dataset.

The algorithm for Naive-Bayes classifier is as follows.

- Initially, the training dataset is passed as input to the algorithm.
- The algorithm calculates the probability of every outcome in the dataset.
- Then, we calculate the probability of every attribute for each outcome in the training dataset.
- During the input of testing data, the calculated values of the attributes in the testing data are used to predict all outcomes.

The outcome with the highest probability value is the predicted outcome for the user's testing data.

K Medoids

This algorithm finds the initial cluster center for obtaining the similar users. The dataset is given as input and the number of clusters to be formed is entered. Based on the number of clusters specified, it iteratively finds the cost for each cluster center. Finally the cluster center with minimum cost is chosen as initial cluster center which is passed on to K-Means algorithm. The algorithm for K Medoids is as follows:

- The dataset is passed as an input parameter to the algorithm.
- The number of clusters required is also given as an input to the algorithm
- The algorithm then calculates the distance between each and every member in the input dataset.
- The distance is calculated using the Manhattan distance formula.

$$d = |X1-X2|+|Y1-Y2|$$

- The two cluster centers with the least distance are the final outcome of K Medoids algorithm.

K Means

This algorithm uses the initial cluster center which is obtained from K Medoids algorithm. It calculates the distance between the current user and all other users in the dataset using Manhattan distance. Thus, the clusters are obtained which specifies the similar users surrounding the cluster center. One of the disadvantages of K-Means algorithm is that it does not find the similarity measure between two users. The algorithm for K Means is as follows:

- The dataset is passed as an input parameter to the algorithm.
- The two users from the K Medoids algorithm are taken as initial cluster centre.
- For each initial cluster centre, do the following:
- Calculate the distance between the initial cluster and other users in the dataset.

$$\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$$

- Compare the distance of every user between the two initial clusters.
- The lesser distance in the first cluster centre forms one cluster and lesser distance in the second cluster centre forms the second cluster. These clusters form the final outcome of K Means algorithm.

Pearson Correlation Coefficient

ISSN: 0975-8585

This method finds the similarity measure between two users.
The following are the possible outcomes of Pearson Correlation Coefficient

- If the coefficient value is less than zero, then the users are negatively correlated.
- If the value is zero, then the users are independent.
- If the value is greater than zero, then the users are positively correlated.

Pearson correlation coefficient can be measured using the following formula:

$$r=\sum xy/(\sqrt{\sum x^2 y^2})$$

Where
r –Pearson Correlation Coefficient
x – Current user
y – Similar user

Term Frequency- Inverse Document Frequency

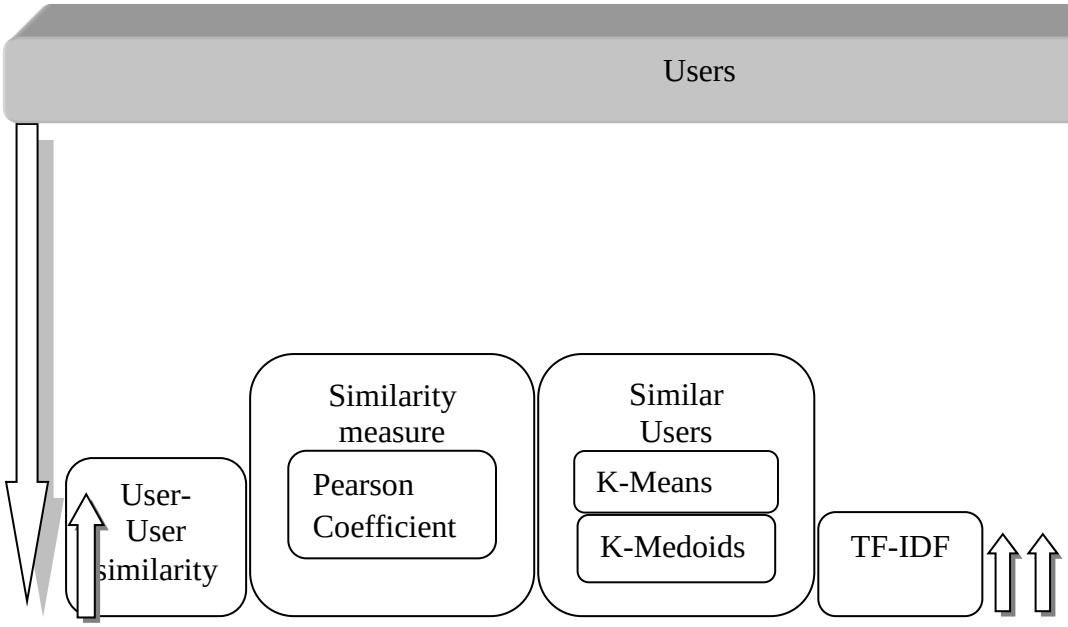
Term frequency is the frequency of a word in a document whereas IDF is the inverse of document frequency among the corpus of documents. The algorithm for TF-IDF is as follows:

- The Term frequency for all attributes is calculated.
- The weight of all attributes is then calculated using the formula

$$w=1+\log(tf)$$

- The length of vector is then calculated by taking the square root of sum of squares of term frequency of attributes.
- The similarity between each and every attribute is calculated using Cosine Similarity. Those attributes with the highest cosine similarity values are taken as output for Term Frequency-Inverse Document Frequency.

The overall workflow of context-Aware mobile search is shown in Fig 2



ISSN: 0975-8585

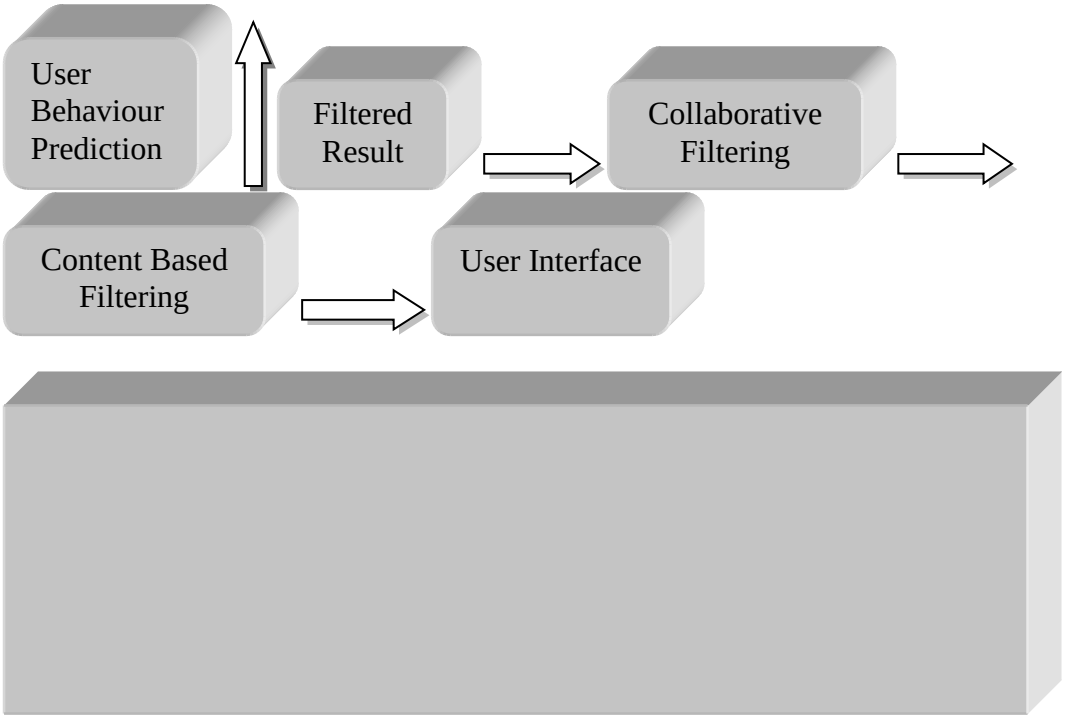


Figure 2: Overall, workflow of context-aware mobile search

EXPERIMENTAL EVALUATION

Data set

We have used 3 kinds of dataset from UCI machine learning repository

They include

- Behavioural dataset (includes behavioural attributes such as budget, rating, comfort, outcome)
- User 's demographic information(includes latitude, longitude measure for identifying user's location)
- User's query information (includes user id, query, link, rating of a link ,number of visit for a link)

To achieve personalized result, initially we have predicted user's behaviour by using Naive-Bayes classifier. To predict user's behaviour, we have used data set which consists of 200 records. Among 200 records, we have used 60% data as training data and remaining 40% data as testing data.

In Table 1, we have shown a sample training dataset for user's behaviour prediction for 5 records

Table 1 : Training set

Budget	Rating	Comfort	Outcome
Medium	2	Business	Class2
Low	2	Economy	Class 4
Low	4	Udeluxe	Class 2
Medium	3	Deluxe	Class 1
Medium	1	Economy	Class 2

The following is the sample testing data set for user's behavioural prediction

Budget: Low
Rating: 1

Comfort: deluxe

From the test data, this classifier uses probabilistic values of each behavioural attributes to predict the user's behaviour. The maximum probabilistic value is considered as outcome for the test data.

P(X/class4) = 0.013084818702567533
X represents the test data.

Among the probabilistic values obtained, it is found that Test data correspond to class 4 since it pertains to maximum probabilistic value.

Likewise we have determined the probabilistic value of remaining test data and predicted user's behaviour. Now we are generating the user model for the predicted users. It is done by extracting some users based on constraint (ie. Retrieve users who belong to class 4 type).

Thereby we are performing filtering for those users who belong to class 4 type.

Content based filtering:

This method filters the link based on user's query history or previous experience. It uses Term Frequency– Inverse Document Frequency (TF-IDF) to find the most relevant links.

A sample user query dataset for user 17 and for query "Hotels in thanjavur" is shown in Table 2

Table 2: User Query Dataset

Link	Content	Num of visit
L1	0.437	0.517
L2	0.812	0.096
L3	0.289	0.333
L4	0.885	0.687
L5	0.975	0.528
L6	0.524	0.726

This table uses attributes such as content, number of visit for links and these attributes belong to particular query and for particular user.

Now the weights of each attribute is calculated by using the formula

1+ log (term frequency)

Length vector is also calculated by taking square root of sum of squares of each attribute and is shown in Table 3.

Table 3 : Sum of Squares for each attribute

Link	Content	Num of Visit	Length vector
L1	1.362557607096888	1.416734700366395	1.9656297830160836
L2	1.594431207620786	1.17898265552844	1.9829803271520308
L3	1.253866723957050	1.2874320411965716	1.7971263233691606
L4	1.6339278208999741	1.5229518035638314	2.2336298529275758
L5	1.6805683983530852	1.4239596907443288	2.2027190793216773
L6	1.4213384572644545	1.5458065926612363	2.099933578000466

We have shown normalized vectors for each attribute in Table 4. This is computed by dividing each attribute to length vectors (as shown in Table 4)

Table 4 : Normalized vectors

Link	Normalized Content	Normalized num visit
L1	0.73106	0.77176
L2	0.913776	0.60313
L3	0.717256	0.74173
L4	0.843918	0.75527
L5	0.896619	0.69368
L6	0.725737	0.82193

After computing the normalized vector for each attribute, we find cosine similarity for each link and obtain the most relevant links.

Cosine similarity is computed in pairs for all combination of links. The dot products for each attribute pairs are computed and are added. The top three pairs of links which have maximum values are considered to be the result of content based filtering. The following is the sample computation shown for content based filtering.

Cos (L1, L2) = 0.73106*0.913776 + 0.77176*0.60313 =1.133497

Similarly the cosine similarity is computed for all combinations of links. The top three maximum link values are taken and these links are passed as input to the collaborative filtering.

Collaborative filtering

This method involves finding similar users to suggest link to the current user. To find the similar users, we have used clustering algorithms such as K-Medoids and K Means. K-Medoids algorithm uses user's demographic information data set (latitude, longitude measure) to finds the initial cluster center. It finds the initial cluster center by computing the cost of all combinations of clusters for a given K value. The least cost value among this combination is considered to be optimal cluster center. In Table 5, we have shown the sample data set that includes user's locational attributes (latitude, longitude).

Table 5 : Sample Dataset includes location attributes

User Id	Latitude	Longitude
1	22.14	-100.979
2	22.15009	-100.983
3	22.11985	-100.947
4	18.867	-99.183
5	22.18348	-100.96
6	22.15	-100.983

In Table 5 we have shown a dataset with 6 records (i e. 6 users). But we have used a dataset comprising of 58 users to find initial cluster center.

From the data set we have found an initial cluster center as (19, 20) for k=2 value using K-Medoids algorithm.

Now we use this initial cluster center as input for K-Means algorithm.
In K-Means algorithm we find the similar users around 19 and 20.
Table 6 shows similar users around user 19 and 20

ISSN: 0975-8585

Table 6: List of Similar Users

Cluster center	Similar users
19	1,2,3,5,6,7,8,9,10,11,13,14,15,16,18,21,22,23,24,25,26,27,28,29,31, 32,33,34,36,37,38,39,43,45,46,47,48,49,50,52,53,54,55,56,57,58
20	4,12,17,20,30,35,40,41,42,44,51

Similarity measure using Pearson Correlation Coefficient

We have used Pearson Correlation Coefficieint to find the similarity measure between similar users and the current user. From the result of correlation coefficient, we have taken users who have positive correlation coefficient. And these users are considered to be the most similar and are allowed for suggesting the link to the current user.

In Table 7 we have shown about User-User similarity to suggest links to the current user.
Let the current user's user id be 17
The query selected by the current user be HOTELS IN THANJAVUR

Table 7: User-User similarity

Similarusers id	Similarity Value	Link1 Rating	sim.link1	Link2 Rating	Sim.link2	Link3 rating	Sim.link3	Link4rating	Sim.link4
4	0.8	1	0.8	2	1.6	2	1.6	4	3.2
12	0.4	3	1.2	4	1.6	1	0.4	1	0.4
20	0.7745	2	1.549	1	0.7745	1	0.7745	2	1.549
30	0.6325	1	0.6325	2	1.265	2	1.265	3	1.8975
Total	2.607		4.1815		4.2395		4.0395		7.0465

Total/sum of similarity for link 1=1.6039
Total/sum of similarity for link 2=1.6261
Total/sum of similarity for link 3=1.5494
Total/sum of similarity for link 4=2.7029

Thus among the 4 values, the top two values which are maximum are taken and those links are considered as outcome of the filtering. And these links are suggested to the users.

PERFORMANCE EVALUATION

We have shown the generic (i.e non personalized result), content based filtering and the Hybrid Filtering results separately (see Fig 3). This is done especially to specify the difference between both types of filtering approaches.

ISSN: 0975-8585

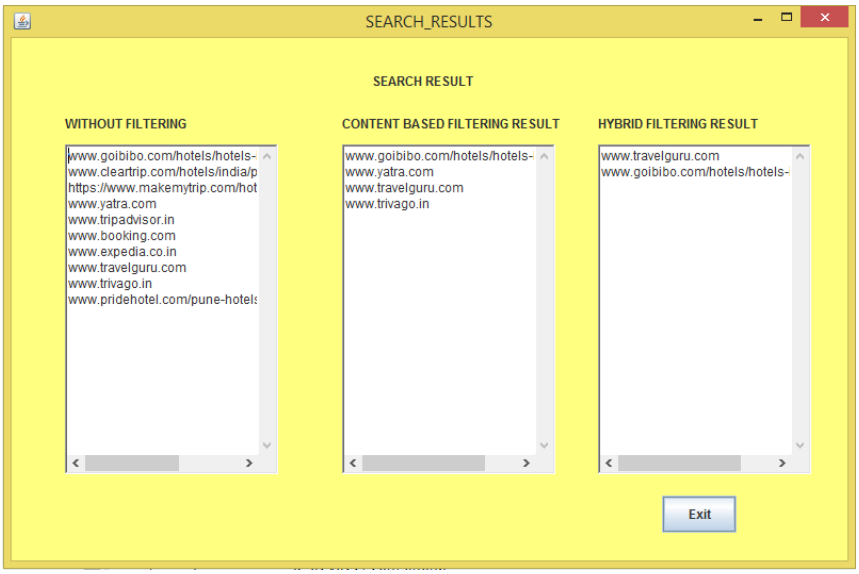


Figure 3: Filtering Results

CONCLUSION AND FUTURE WORK

We have suggested a context-aware search approach called as Hybrid filtering approach to filter links based on user's behavior. Initially we predicted user's behaviour to enhance the effectiveness of the hybrid filtering approach. Then, we have used K Medoids algorithm which gives the initial cluster centre as output. And this initial cluster center is passed to K-Means algorithm instead of a random cluster centre. We have also found the similarity measure between the current user and each of the similar users using Pearson Correlation Coefficient. By finding similarity measure we ensure that the similar users who are having a positive similarity measure are able to suggest links to the current user. Therefore, we infer that the K Medoids algorithm improves the performance of the clustering algorithm. Thus the hybrid filtering approach will enhance the personalized mobile search experience significantly.

Currently, we have implemented the filtering approach for specific queries and we have designed a framework wherein we have included few applications only. So in future, we are planning to include more applications in our framework so that filtering can be done on different queries. Secondly, we plan to perform semantic analysis on queries so that limitation of filtering on specific queries can be avoided and filtering can be performed on generic query also.

REFERENCES

[1] Krzyszt Dembczynski, Wojciech Kotłowski, Marcin Sydow,Effective Prediction of Web User Behaviour with User-Level Models,Fundamenta Informaticae 89,2008, IOS Press, 1–18.

[2] Santra.A.K, Jayasudha.S,Classification of Web Log Data to Identify Interested Users Using Naive Bayesian Classification,International Journal of Computer Science Issues,2012, 9(1) .

[3] Supreet Dhillon, Kamaljit Kaur,Comparative Study of Classification Algorithms for Web Usage Mining, International Journal of Advanced Research in Computer Science and Software Engineering, July 2014, 4(7).

[4] Arne Mauser, Ilja Bezrukov, Thomas Deselaers, Daniel Keysers Lehrstuhl f¨ur Informatik VI,Predicting Customer Beha vior using Naive Bayes and Maximum Entropy, Computer Science Department RWTH Aachen University, Germany.

[5] Masud Karim, Rashedur M. Rahman, Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing,Journal of Software Engineering and Applications, 2013, 6, 196-206.

[6] Simon Philip, Shola.P.B, Abari Ovye John,Application of Content-Based Approach in Research Paper Recommendation System for a Digital Library,International Journal of Advanced Computer Science and Applications,2014,5(10).

[7] Joonseok Lee, Mingxuan Sun, Guy Lebanon, A Comparative Study of Collaborative Filtering Algorithms 2012.
[8] Michael J. Pazzani, A Framework for Collaborative, Content-Based and Demographic Filtering, Artificial Intelligence Review,1999.
[9] Torres, Roberto, Combining collaborative and content-based filtering to recommend research papers ,Porto Alegre,2004.
[10] Poonam B. Thorat,Goudar R.M, Sunita Barve,Survey on Collaborative Filtering, Content-based Filtering and Hybrid Recommendation System,International Journal of Computer Applications,2015,110(4).
[11] McCallum, Andrew, Kamal Nigam, and Lyle H. Ungar,Efficient clustering of high-dimensional data sets with application to reference matching,Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining,2000.
[12] Sankar Rajagopal, Customer Data Clustering Using Data Mining Technique,International Journal of Database Management Systems , November 2011,3(4).
[13] Amandeep Kaur Mann, Navneet Kaur, Survey Paper on Clustering Techniques, International Journal of Science Engineering and Technology Research,April 2013, 2(4).
[14] Jadhav Bhushan G, Warke Pushkar U , Kuchekar Shivaji P, Kadam Nikhil, Searching Research Papers Using Clustering and Text Mining,International Journal of Emerging Technology and Advanced Engineering.
[15] Rakesh Chandra Balabantaray, Chandrali Sarma, Monica Jha, Document Clustering using K-Means and K-Medoids,International Journal of Advanced Research in Computer Science and Software Engineering.