A Personalized Online News Recommendation System

Saranya.K.G
Department of Computer Science and Engineering,
P.S.G College of Technology
Coimbatore, TamilNadu- 641004, India

G.Sudha Sadhasivam, PhD.
Department of Computer Science and Engineering,
P.S.G College of Technology
Coimbatore, TamilNadu- 641004, India

ABSTRACT

Traditional news recommendation systems strive to adapt their services to individual users by virtue of both user and news context information. This paper describes personalized news recommendation approach based on dynamic updating policy and collaborative filtering. Adaptive user profiling is a principled framework for news selection based on the intrinsic property of user interest presented, with a good balance between the novelty and diversity of the recommendation result. Also it considers the exclusive characteristics like news context, access patterns, popularity of the news and recency. Collaborative filtering approach can efficiently capture user's behavior in case where the overlap in historical assumptions across users in relatively high and the context universe is almost static. The major issue with the personalized news recommendation system is scalability. This paper addresses the above mentioned issue with the help of hadoop framework. Experiments on a collection of sports related news obtained from various news websites demonstrate the efficiency of the proposed approach.

Keywords: Adaptive User Profiling, Dynamic Updating Policy, Collaborative Filtering

1. INTRODUCTION

World Wide Web has made it possible for people to access and generate electronic information easily. With the advances and proliferation of the internet, available information sources have grown tremendously in number and sheer volume, primarily as a result of global connectivity and ease of publishing. Traditional information retrieval techniques often follow the one-size-fits all policy by delivering the queued information in the same form and order for every user with the same query. Recommendation systems which have emerged in response to the above challenges provide users with recommendation of content suited to their needs. Since different user information needs and queries arise in varying contexts with different intentions, research has started to focus on delivering tailored, adapted and personalized information presentations. A challenging problem is how to efficiently select specific news articles from a large corpus of newly published press releases to individual readers, where the selected news items should match the reader's reading preference. This is referred as personalized news recommendation.

Personalized news recommendation requires, content in machine readable format, Knowledge of user's requirements and reasonable query execution time. In this paper, personalization task in news recommendation system has been achieved by combing adaptive user profiling and collaborative filtering technique. Adaptive user profiling using dynamic updating policy considers the change of the user's requirements over time and domain. The traditional task in collaborative filtering is to predict the utility of a certain item for the target user from the opinion of other similar users, and thereby make appropriate predictions. This paper presents three specific contributions towards user modeling in news recommendation systems: it includes, building a user model, allowing, understanding and filtering of the user's interest and application of the model to personalized recommendations relevant to the user's needs. 3. Incorporation of collaborative filtering characteristics into an adaptive user profiling using dynamic updating policy.

2. RELATED WORKS

The major challenge within the traditional information retrieval paradigm is to match a query with documents and rank documents according to their relevance value. As a result the whole information retrieval process is an independent cycle of query submission and result display, which is inadequate for exploiting user context (X. Shen,2007). (Fikadu Gemechu and Zhang Yu, Liu Ting ,2010)proposed a framework in which a user profile module is incorporated as an integral component of an information retrieval process so that results returned by traditional retrieval system are filtered based on the user's profile to meet users' specific information need.(Hochul Jeon, Taehwan Kim and Joongmin Choi,2008) proposed a model for personalized information retrieval. The user's information based on user's information needs should be extracted to provide users with personalized information.The user's model of information needs should be established by calculating users' similarity. The information should be recommended to users by similar user groups. In this way, users can be provided with potential information to meet their individual needs. Information recommending via utilizing user interests can commonly be divided into content-based recommending and collaborative recommending (J. Ben Schafer et al.., 2007).

The system describe by (J. Wang et al..,, 2006)focuses on the change of user preferences. The core problem of personalized recommendation is to model and track users' interests and their changes. To address this problem, both content-based filtering (CBF) and collaborative filtering (CF) have been explored in this system. User interests involve the interests on fixed categories and dynamic events, yet in the current CBF approaches, there is a lack of ability to model user's interests at the event level. So, the system (J. Wang, Z. Li et al..,2006) proposed a novel approach to user profile modeling. In this model, user's interests are modeled by a multi-layer tree with a dynamically changeable structure, the top layers of which are used to model user interests on fixed categories, and the bottom layers are for dynamic events. Thus, this model can track the user's reading behaviors on both fixed categories and dynamic events, and consequently capture the interest changes. (H. Naderi, and B. Rumpler, 2006) uses both query based and document based approaches to partially capture the users' interests. Collaborative filtering technique suffers from a cold start problem. The cold start problem is most prevalent in recommender systems. Recommender systems form a specific type of information filtering technique that attempts to present information items (movies, music, books, news, images, web pages) that are likely of interest to the user. The cold-start problem occurs when it is not possible to make reliable recommendations due to an initial lack of ratings of a particular community.. The proposed news recommender system adopt a hybrid approach between content-based matching and collaborative filtering. Content-based filtering technique try to sequentially find newly-published articles similar to the user's reading history in terms of content.

3. SYSTEM IMPLEMENTATION

3.1. Issues with Personalized news recommendation System

In general, people look for instantaneous access to fresh news events. When surfing on the internet, user looks for interesting news events among a large amount of news articles. Typically, news reading services retrieve news articles relevant to reading preferences of individual users, and adapt their services based on the change of the user's reading interests by employing different recommendation approaches.

3.1.1 Unique characteristics of news items are as follows:

1. Large volume: Unlike other types of web objects, news articles tend to be in flood within a short period of time, requiring much more computation for recommendation.

2. Unstructured format: The unstructured format of a news story is more difficult to analyze than other objects with structured properties.

3. Recency: News items typically have short self lives. For example, few sports fans will be concerned with the past breaking scores of two days ago. In contrast, the shelf lives of products and movies extend several months or even years.

4. Entity preference: Most news articles describe the occurring of specific events. Online news readers tend to be interested in the named entities of information like what, when, who and where the event occurred.

5. News selection and Ranking: The interest of news articles with respect to a user is regressive. i.e.., after he/she clicks the first piece of news he/she interested in, the interest of the user on a news item may change if he chooses other news items also.

6. Scalability: The scalability of news recommendation requires elegant algorithms to efficiently deal with large news corpus. Map-reduce, a programming model prototype by Google, aims to support distributed computing on large datasets with clusters of computers and has been widely used in many data mining and

   Machine learning tasks.

7. Large datasets with clusters of computers and has been widely used in many data mining and machine learning tasks.

The hadoop framework has been used to efficiently handle the above mentioned issues.

Hadoop is an open source distributed parallel computing platform and it has the reliability, efficiency and scalability. It mainly consists of a parallel computing framework Map-Reduce and a distributed file system HDFS. Large volume of data's are stored and managed in HBASE [8] is a column oriented database. The data in an HBASE instance is laid out more like a Hash table, and the data is immutable. HBASE is an open source; non-relational, distributed database modeled after Google's big table and is written in java. With the help of HBASE, data can be de-normalized. HBASE is good for semi-structured data as well as structured data. It provides fault tolerant way of storing large quantities of sparse data [9].

3.1.2 Proposed News Recommendation Framework

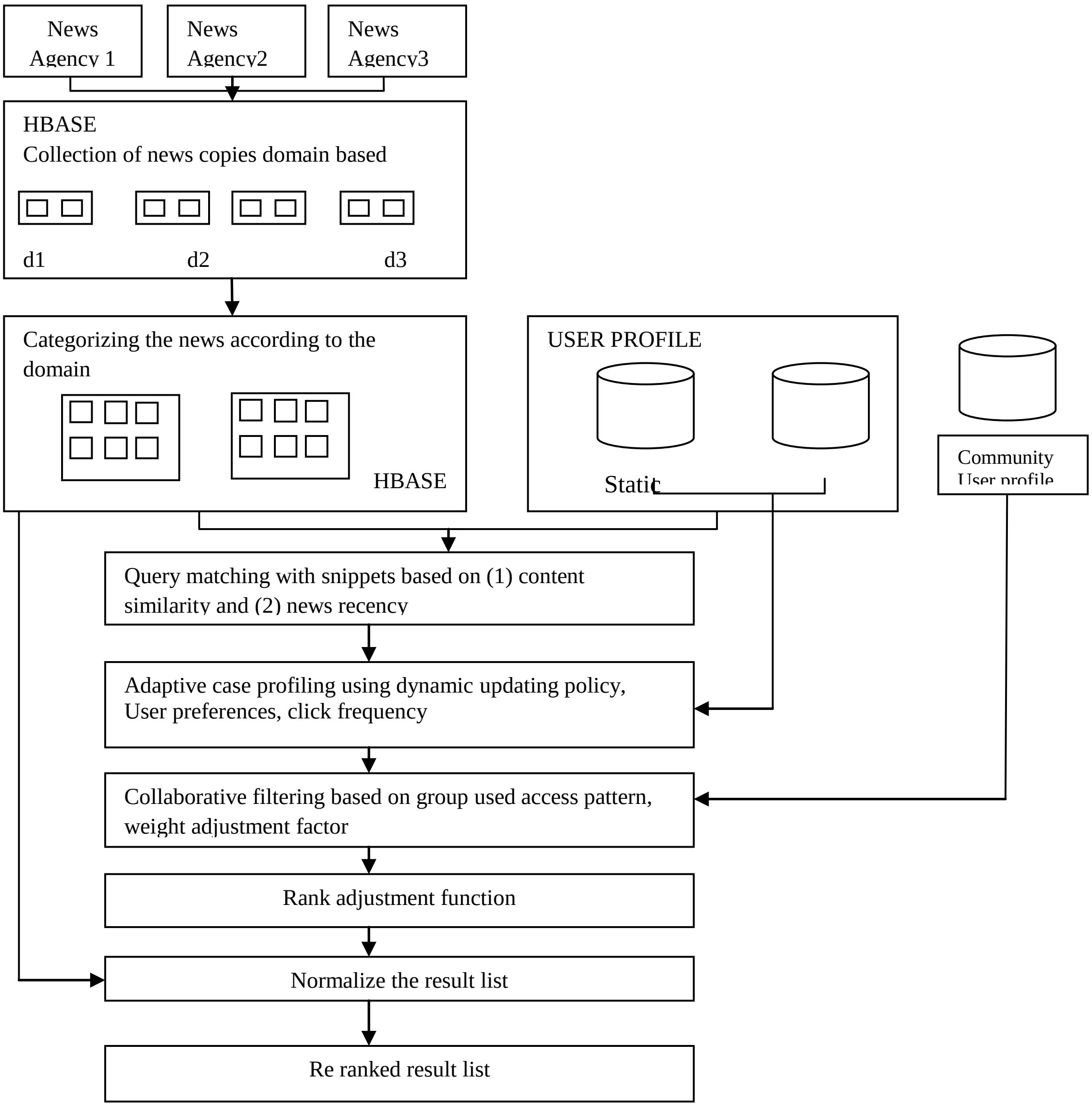

News Agency 1
News Agency2
News Agency3
HBASE
Collection of news copies domain based
d1
d2
d3
Categorizing the news according to the domain
HBASE
USER PROFILE
Static
Community User profile
Query matching with snippets based on (1) content similarity and (2) news recency
Adaptive case profiling using dynamic updating policy, User preferences, click frequency
Collaborative filtering based on group used access pattern, weight adjustment factor
Rank adjustment function
Normalize the result list
Re ranked result list

Fig.1. Proposed News Recommendation Framework

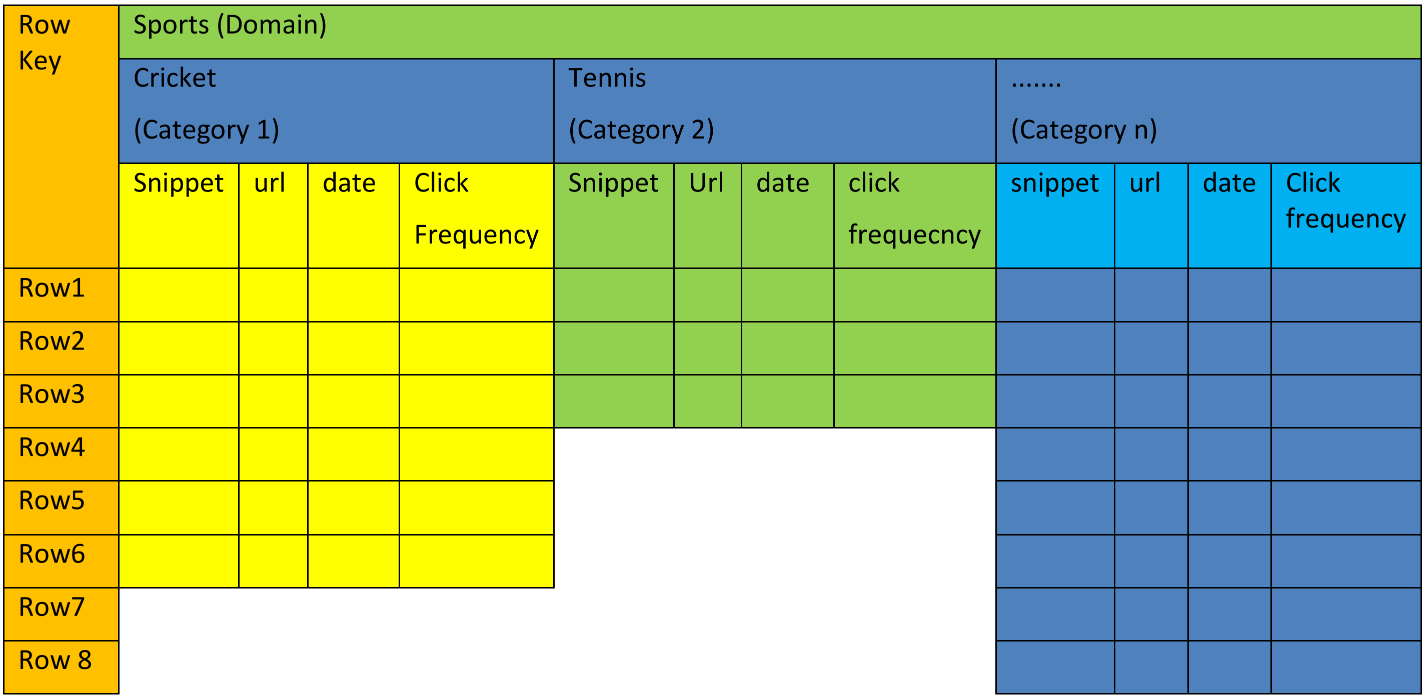
Various steps as in figure.1 is discussed as follows:

3.1.2.1. Original News Corpus:

Newly published news corpus is collected from different new's agencies through web and dump it into the HBASE on a daily basis. In the proposed framework, last one month news will be present. Whenever newly published news collections are dumped into the HBASE, all the out dated news will be automatically deleted.

3.1.2.2. Categorizing the News articles:

News articles are clustered into small groups based on its category. A Map-Reduce framework is used to match every document with its associated category. Newly published news collection's are categorized into small groups and stored in the HBASE as snippet data base. And it is purely based on news content. The snippet data base is as follows.



Row Key	Sports (Domain)											
	Cricket (Category 1)				Tennis (Category 2)				.... (Category n)			
	Snippet	url	date	Click Frequency	Snippet	Url	date	click frequecncy	snippet	url	date	Click frequency
Row1												
Row2												
Row3												
Row4												
Row5												
Row6												
Row7												
Row 8												

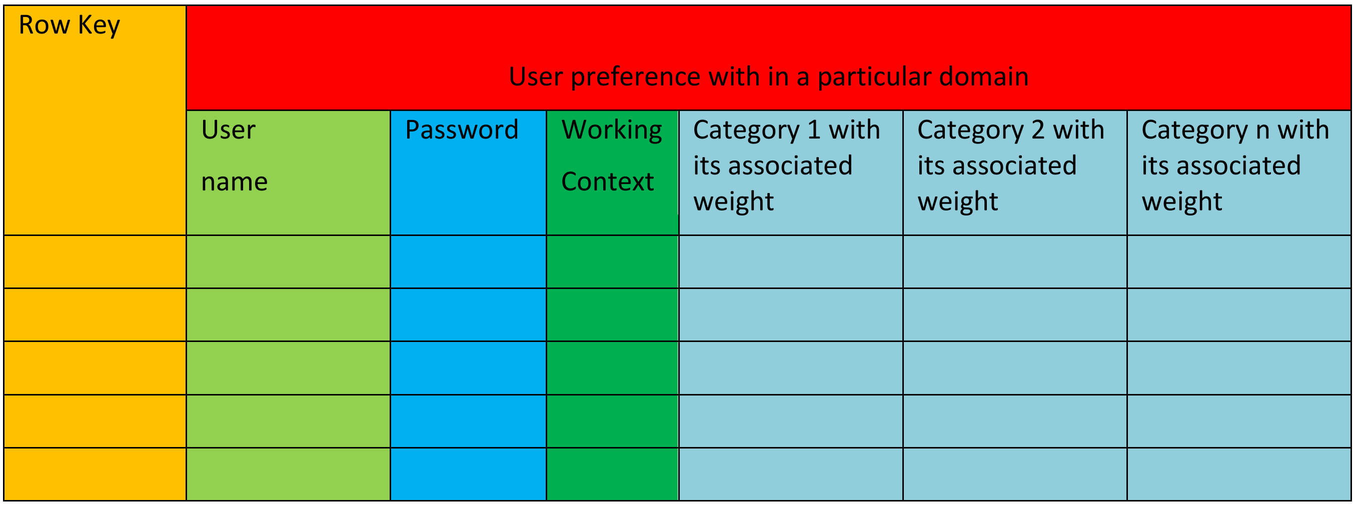
Fig 2. Hbase Structure of Snippet Database

3.1.2.3. Query Matching With Snippets:

User query's are matched with the snippets present in the snippet data base based on the content similarity and news recency.

3.1.2.4. User Profile Construction:

A user profile can be defined by keeping track of the history of user's interest. In the proposed approach, two types of user profiles are constructed. One is static user profile and another one is dynamic user profile. Static user profiles are constructed using information gathered from the user during sign up. It contains the information like user name, password, working context, favorite/hobby. Dynamic user profile is used to address frequently changed user interest. The explicit interest indicators enable the content-based filter to use direct learning in predicting news category interest. Dynamic user profiles are constructed during every interactive search session initiated by particular user. It contains the information like user name, password, news content (accessed), click frequency for each news content, similar access pattern, weight associated with each document belonging to particular category. All these factors are extracted from the user's reading history. The implicit interest indicator enables the content-based filter to use indirect learning in predicting news category interest. Initially weights for each category is assigned a rating scale between 0 and 1, based on user interest gathered explicitly during sign up. Sum of the weight of all categories with in a particular domain will be 1. And this weight will be updated every time after the completion of the particular search session. Since user preferences will be changed over time, the proposed system gives an opportunity to make modifications in the static profile every time after the completion of the particular interactive search session. Here news content is denoted as a topic distribution of the reading history. Similar access patterns are generated by analyzing click behaviors of different users.



Row Key	User preference with in a particular domain					
	User name	Password	Working Context	Category 1 with its associated weight	Category 2 with its associated weight	Category n with its associated weight

Fig.3. Hbase structure of user profile data base

3.1.2.5. Personalized Recommendation:

Personalized news recommendation is made by calculating the similarity between topic distributions of each news group and the user's reading history using adaptive user profiling and collaborative filtering techniques. By interacting with the sign up frame, user can log on to the system. User can post their query through a recommendation frame. Main frame will provide a personalized search results and it will update the user's dynamic profile. By interacting with the favorite frame user can change his/her interest explicitly.

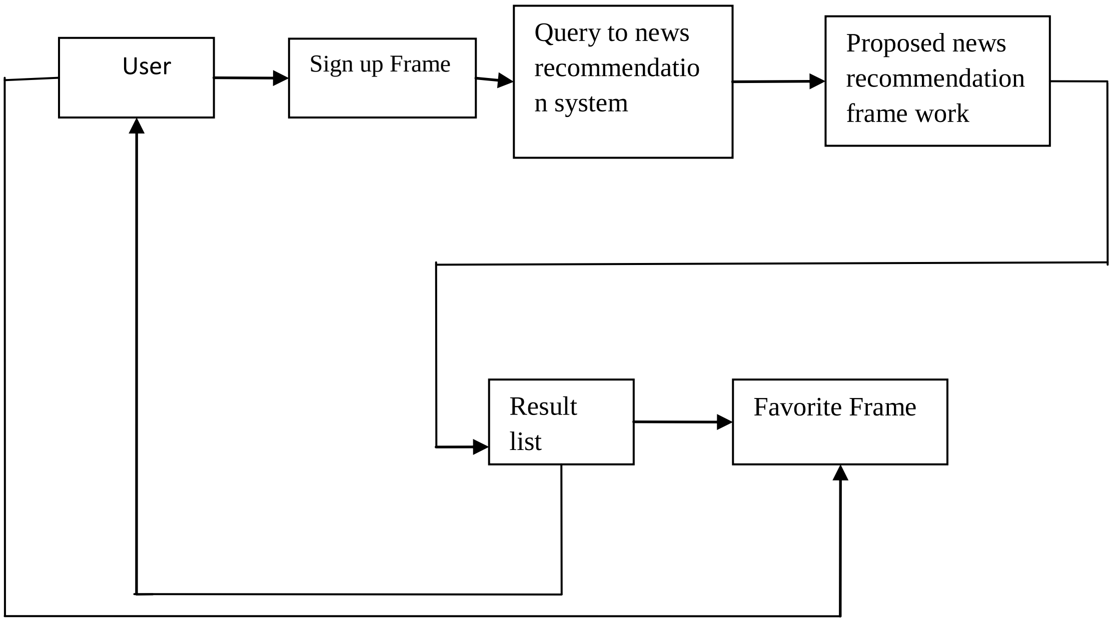

User
Sign up Frame
Query to news recommendation system
Proposed news recommendation frame work
Result list
Favorite Frame

Fig 4. Personalized News Recommendation Architecture

3.1.2.6 Adaptive User Profiling Using Dynamic Updating Policy:

For a newly registered user, system will analyze user registration information. Based on the user's preference given at the time of sign up and the other user preferences that belongs to the same preference category will be taken by the system. Whenever a user clicks a particular document, its click frequency is updated by 1. After a particular search session has been completed by a particular user, user's profile information is updated. Category weight in user's profile is updated using equation (2). If a registered user signs up into the system, the search result will be based on his/her preference stored in the user's profile and the community weight associated with the particular category of interest. If a user visits a news item belonging to a category, not in the user profile, then the newly visited document and its associated category will be updated in the user profile. Simultaneously, the weight and the click frequency of the item is updated. Every non clicked document's click frequency is considered as 0. The function f denotes the resultant document (d) list for a particular query belongs to particular category (c_j) submitted by the user (a) based on user profile.

$$f(a, c_j, c_d, cf) \quad = W_{a,c_j,c_{new}(wt)} \times W_{a,c_d,cf} \tag{1}$$

× - represents multiplication

$$C_{new}(wt) = \frac{W_{a,c_j,c_{old}(wt)}}{\sum_{j=1}^{n} W_{a,c_j,c_{old}(wt)} \times W_{a,c_d,cf}} \tag{2}$$

Where,

a=active user

u=user

$c_{new}(wt)$= weight updation factor

$cf =$ Click frequency

$c_j =$ Query category

$c_{old}(wt) =$The value of weight computed in previous iteration using weight updating function

$c_d =$News item belong to particular category

3.1.2.7. Computing Dynamic Neighborhood (Collaborative Filtering Technique):

The dynamic neighborhood of the user is computed with respect to the query category. It involves two steps: at first, all the users having sub profiles in the query category are retrieved. Second, for each of this users having sub profiles in the category, the proposed approach computes his/her similarity with the active user using the equation (3). The proposed approach assumes that the category of the query is given. The function f denotes the similarity between a user u whose sub profile in this category (c_j) with its associated weight(f^n(wt)) and click frequency(cf) is $w_{u,}$ c_j and the active user(a) whose sub profile in this category *is* $w_{a,}$ c_j,

$$f(a, u, c_d, c_j, cf) = 2W_{a,c_j} \frac{\left[\frac{\sum_{u=1}^{n}(W_{a,c_j,c_{new}(wt)} \times \sum_{d=1}^{n} W_{u,c_d,cf}) + (W_{a,c_j,c_{new}(wt)} \times \sum_{d=1}^{n} W_{a,c_d,cf})}{2}\right]}{W_{a,c_j} \times \sum_{i=1}^{n} u_i} \tag{3}$$

Where, X- Denotes Multiplication. Then the users are sorted down based on the f value and picked the top k users.

3.1.2.8. Calculating Rank of a Document:

The rank of a document is computed based on the rank for a document computed with respect to the active user and the community rank for the document using the equation (4).The rank for a document with respect to the active user is computed using f described in equation 1. The community rank of a document is the average rank computed with respect to all the users in the computed neighborhood weighted by the similarity between the given user and the user in the neighborhood computed using f described in equation(3).

The rank of a document d for the query q belongs to particular category c_j with respect to the user (a) is calculated as,

$$R_{a,d,c_j} = \left(W_{a,c_j,c_{new}(wt)} \cdot W_{a,c_d,cf}\right) + \sum_{top\ k\ uses} f(a, u, c_d, c_j, cf) + default\ result\ list\ from\ recomendation\ system \tag{4}$$

"+" denotes concatenation

Finally, the resultant documents are normalized and displayed according to the rank associated with each document.

4. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed system, sports column in different e-news papers were taken. 500 news items were taken for each query under sports domain. In this experiment 26 members were presented in a group. The results of 25 different query's belongs to different categories like cricket, tennis, football, athletics, hockey under the sports domain were taken. the results from Indian e-news papers the Hindu, India times, Indian express, the times of India were compared against with the proposed system. F-Measure @20 is used to evaluate the performance, the most widely used metrics for evaluating approaches performing re ranking of results. It measures the number of relevant documents found in the top 20 results. Top 20 results from all the news papers were taken to calculate the F-measure (f-measure @20). The results were shown in the table1, and the graphical results were obtained in the figure.5. The higher the f-measure value indicates that the more relevant results are retrieved from the proposed news recommendation framework. The accuracy of the proposed approach is shown in the table.2, and the graphical results of the accuracy were shown in the figure.6.

Table1. Performance Comparison of Different News Recommendation Systems against Proposed System

Various News websites	User Query belongs to Five Different Category				
	Cricket	Tennis	Football	Athletics	Hockey
	F-Measure@20	F-Measure@20	F-Measure@20	F-Measure@20	F-Measure@20
THE HINDU	0.098	0.049	0.063	0.022	0.045
TIMES OF INDIA	0.07	0.05	0.073	0.012	0.077
INDIAN EXPRESS	0.085	0.078	0.029	0.029	0.03
INDIA TIMES	0.058	0.076	0.055	0.054	0.04
PROPOSED SYSTEM	0.133	0.087	0.085	0.07	0.113

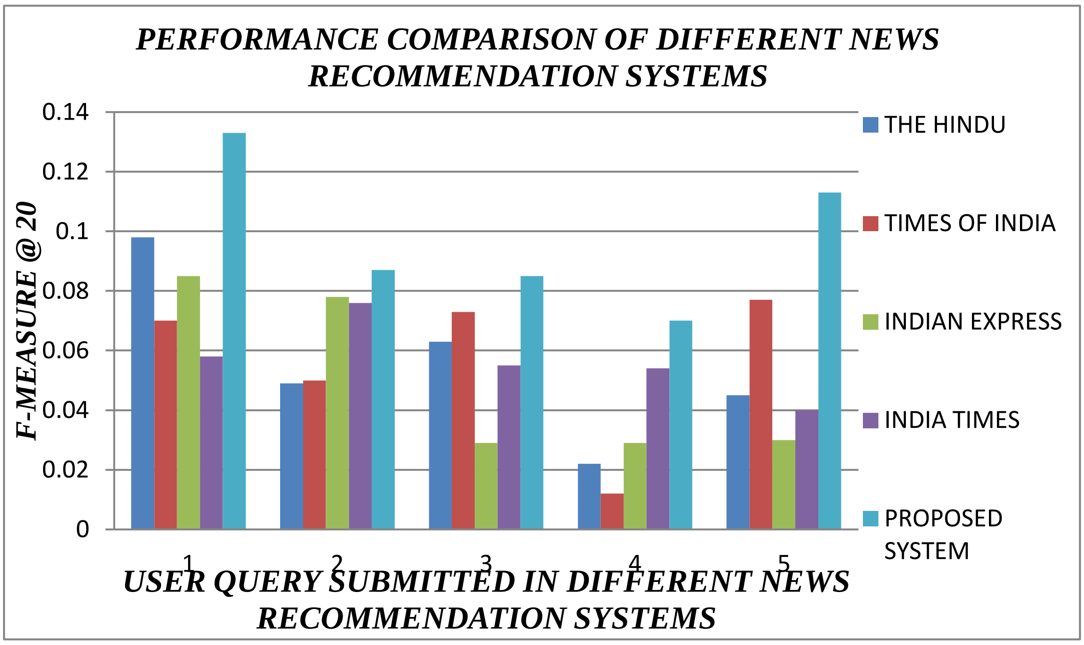


Fig.5. Graphical Representation of the Experimental Result

Table2. Accuracy of Different News Websites

Various News websites	User Query belongs to Five Different Category under Sports domain				
	Cricket	Tennis	Football	Athletics	Hockey
	Accuracy@20	Accuracy@20	Accuracy@20	Accuracy@20	Accuracy@20
THE HINDU	0.65	0.5	0.75	0.5	0.55
TIMES OF INDIA	0.65	0.5	0.5	0.55	0.6
INDIAN EXPRESS	0.5	0.55	0.3	0.4	0.75
HINDUSTAN TIMES	0.45	0.35	0.55	0.3	0.25
PROPOSED SYSTEM	0.95	0.9	0.95	0.9	0.9

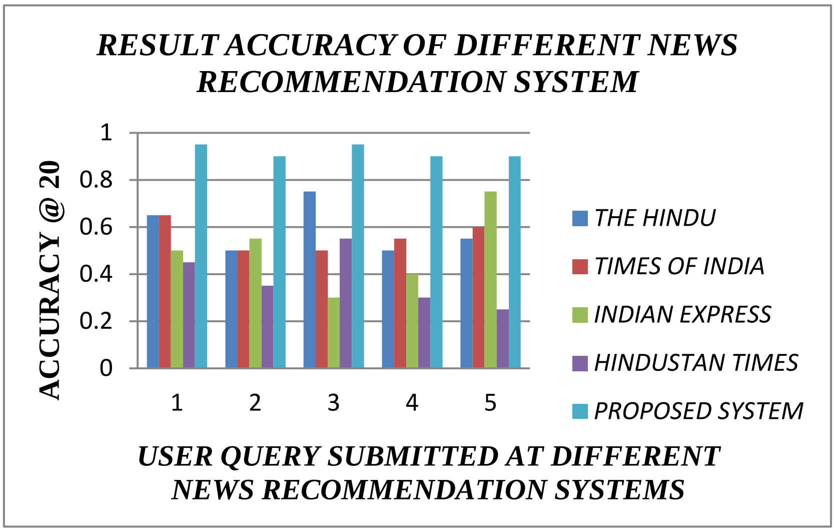


Fig.6. Graphical Representation of the Accuracy of Different News Websites

5. CONCLUSION

Personalized recommender systems are becoming widely used solution for reducing information overload of diverse domains. This paper describes a new and unique method for modeling user interests via a adaptive user profiling and collaborative approaches of different users. It also provides enhanced recommendation accuracy. The major advantage of the proposed modeling method is that it supports not only considers the change of user preferences over time and domain, identification of each user's useful patterns but also enrichment of valuable neighbors' patterns. From the experimental results, it is observed that the proposed method yield better recommendation accuracy compared to the benchmark methods. The proposed method can provide more suitable content for user preferences, even when the number of recommended items is small.

6. ACKNOWLEDGEMENT

The authors like to thank Dr R Rudramoorthy, Principal, PSG College of Technology, Coimbatore, and Mr.Chidambaram Kollegncode, Director, Cloud Computing, Yahoo Research India.Pvt.Ltd., Bangalore for providing us the required facilities and support to complete the project. This project is a collaborative research done by Nokia and PSG College of technology.

7. REFERENCES

[1] X. Shen, “User-centered adaptive Information Retrieval” ,PhD thesis in Computer Science, University of Illinois Urbana-Champaign, 2007.

[2] Fikadu Gemechu, Zhang Yu, Liu Ting, “A Framework for Personalized Information Retrieval Model “, Proc. of IEEE transaction, 2010.

[3] Hochul Jeon, Taehwan Kim, Joongmin Choi,” Adaptive User Profiling for Personalized Information Retrieval”, Third 2008 International Conference on Convergence and Hybrid Information Technology

[4] C. Zeng, C. Xing, and L. Zhou, “A Personalized Search Algorithm by Using Content-Based Filtering”, Journal of Software, 2003,14(5), pp. 999-1004.

[5] J. Wang, Z. Li, J. Yao, Z. Sun, M. Li, and W. Ma, “Adaptive User Profile Model and Collaborative Filtering for Personalized News”, In Proceedings of the APWeb 2006, pp. 474-485, 2006.

[6] H. Naderi, and B. Rumpler, “PERCIRS: a Personalized Collaborative. Information Retrieval System”, In Proceedings of the INFORSID, pp. 113-127, 2006.

[7] J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen,” Collaborative Filtering Recommender Systems” Springer-Verlag Berlin Heidelberg, pp. 291 – 324, 2007

[8] Hadoop Site. http://hadoop.apache.org.

[9] Hadoop Map/Reduce tutorial.