International Journal of Computer Science and Business Informatics

IJCSBI.ORG

Personalizing Education News Articles Using Interest Term and Category Based Recommender Approaches

S. Akhilan
Research Scholar, Department of Computer Science and Engineering
National Institute of Technology, Tiruchirappalli – 620 015.
Tamil Nadu, INDIA.

S. R. Balasundaram
Associate Professor, Department of Computer Applications
National Institute of Technology, Tiruchirappalli – 620 015.
Tamil Nadu, INDIA.

Abstract

With the growth of internet technologies, numerous ways and mechanisms are being developed so that any information to any person about any entity can be delivered irrespective of time and place. Knowing about the happenings around the world is the primary interest of almost every individual. This is achieved with the help of various news service providers such as Yahoo, Google, Your News etc. While delivering news to its users most of the news service providers do not take into the account the user's choice or interests. Providing all news to all will not be an appropriate one when there exists different class of viewers. This major problem can be solved by adopting personalization and it is the key factor which aims at providing the appropriate data for the related person. Personalization in the context of education has resulted in lot of benefits to learners. When recommending news, including educational related, most of the traditional approaches are based on TF-IDF, i.e., a term-based weighting method which is mostly used in information retrieval and text mining. However, many new technologies have been made available since the introduction of TF-IDF. This paper proposes certain new methods for recommending educational news items based on Category Term Based(CTF-IDF) and Weighted Category Term Based (WCTF-IDF) approaches. CTF-IDF is built and also tested in Athena, a recommender extension to the Hermes Genesis News Platform (HGNP). Experiments show that compared to term based, our approaches such as category term and weighted category term based approaches perform better. Also, Athena based recommender provides better results.

Keywords: Term based classification, user profile, educational news items, personalization, category based classification.

1. INTRODUCTION

News is an entity through which an individual can come to know what had happened as well as what is happening around him/her. With the enhanced technologies of World Wide Web, the methods adopted to read news content have changed dramatically from the traditional model of news reading through physical news paper to access millions of web sources via internet. News service providers such as Google News, Yahoo News, etc collect news from various sources and provide an aggregate view of news to the users around the world. The available news documents make the user to feel that they are overloaded with lot of news contents. To overcome this issue, it is a challenging task to find out the users choice of interests in reading the news articles. In response to this challenge, information filtering is a technology that helps the user to retrieve what they need. Based on a profile of user interests and preferences, systems recommend items that may be of interest to the user [1]. Especially, educational news correspond to various items such as knowing about education articles, universities/institutions, courses, reading materials, events etc.

In the present web scenario, recommendation systems play a vital role in delivering the required news to the required users[2]. Content based recommendation is one of the often used recommendation methods. Several content based recommenders for news personalization deploy TF-IDF and cosine similarity measure. Many times a keyword (term) may be useful to extract more number of documents. Combined with vector space model this approach may recommend more news items to any user. In order to obtain news documents pertaining to the related terms of a keyword, an enhancement to TF-IDF is suggested in this paper. We refer to CTF-IDF and WCTF-IDF as classification method that combines the key concepts of term based traditional based classification.

When employing user profiles that describe users' interest based on the previously browsed items or profile, these can be translated into vectors ofTF-IDF weights [3]. Combining the related terms for certain term can be grouped based on concepts or categories. User profiles are used to extract the required terms from the data source based on interest terms. One of the strategies to obtain user terms is through browsing pattern [4]. In this approach user may search or click only specific terms related to his/her areas of interest. For example a user interested in knowing about conferences may browse the keyword ‘conference’. The web portal may

provide results based on conference or few related terms thereby improving the results of the user. The proposed method is tested under Athena, an extension of Hermes Framework.

The structure of the paper is as follows. Related work is discussed in Section 2 followed by methodology in Section 3 and 4. In Section 5, proposed classification method is discussed. Athena framework which is the implementation of Hermes News Portal is discussed in Section 6. In section 7 results of the proposed method is discussed followed by conclusion in Section 8.

2. LITERATURE REVIEW

2.1 News Personalization

Information filtering plays a critical role in recommender systems and thereby in news personalization. It prevents recommending information that are not been rated and accommodates the individual differences between users [5]. Apart from news domain, information filtering is applied to various fields such as email, e-commerce, etc [6]. In the domain of news, this technology particularly aims at aggregating news articles according to user interests and creating a "personalized newspaper" for each user.

Recently, personalized news recommendation has become a desirable feature for websites to improve user satisfaction by tailoring content presentation to suit individual users' needs [7].

2.1 Classification Methods

Personalization involves a process of gathering and storing user attributes, managing content assets, and, based on an analysis of current and past users' behaviour, delivering the individually best content to the present user being served. Personalization can be defined as the use of technology and user information to tailor the web news documents as per the requirements of an individual who wants to access the news articles from different news web sites providers. Each news provider delivers numerous dynamic news updates collected from various sources. While getting news from service providers or news portals, they are delivered with news contents which are un related to the users. To overcome this Buckley et al. (2009) has proposed different aspects of personalization in their system such that users are delivered with required news contents. Cleverdon et al. (2009) proposed that by creating personalized sites where a user can add his/her own interests can view the most recent and popular news. In personalized news classification, users can define their personalized categories using few keywords [8, 9]. As per Mills et al. (2009) personalization includes the attempts that have been

made by the major search engines and portals considers only the issue of viewing already categorized content according to the user's interests. Classification is a methodology to classify documents of varied domains based on a particular interest of a user. There are numerous classification methods as shown in figure 1.

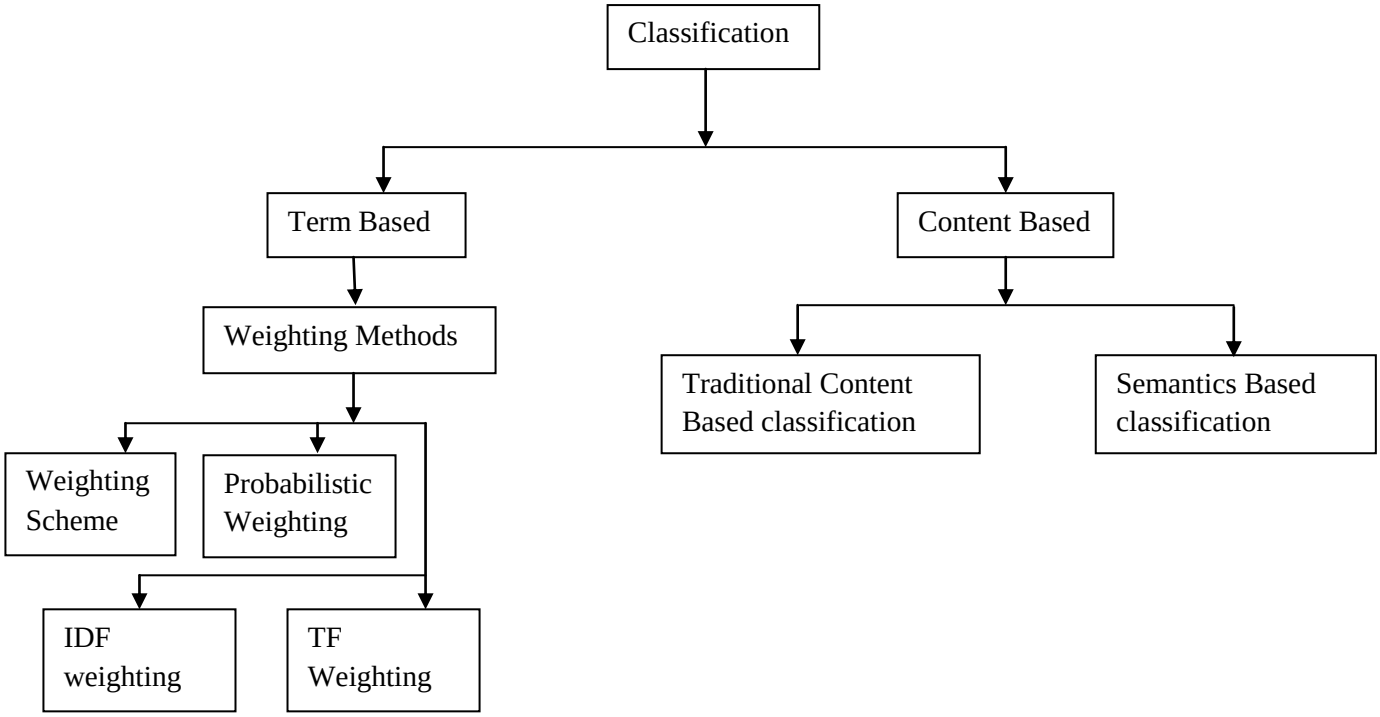

Figure 1: Various Classification Methods

2.2 Term Based Classification

With the advent of web technology it is possible to retrieve any information irrespective of time and place. The content of Web is vast and it is dynamic in nature. This causes the users to feel discomfort in using the Web documents. For a user query, search engines deliver many documents. The documents that are all delivered may not be useful to all users. For this reason, term based classifications methods are adopted to reduce the number of contents delivered to a user such that the user receives the required documents only. In this regard, TF-IDF is the most prominent one used for classifying the terms within documents.

TF-IDF, term frequency – inverse document frequency, is a numerical statistic which reflects how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining [6, 19, 20]. TF-IDF value increases proportionally

to the number of times a word appears in the document. But, it is offset by the frequency of the word in the corpus, which helps to control the fact that some words are generally more common than others. TF – IDF is the product of two statistics, term frequency and inverse document frequency. Various ways for determining the exact values of both statistics exist. In the case of the term frequency tf(t,d), the simplest choice is to use the raw frequency of a term in a document, i.e. the number of times that term t occurs in document d. If we denote the raw frequency of t by f(t,d), then the simple tf scheme is tf(t,d) = f(t,d).Stop words are filtered from the documents before calculating the TF-IDF values. The remaining words are stemmed by a stemmer. Finally the term frequency is calculated which indicates the importance of a term within a document. In figure 2 term based classification approach is shown.

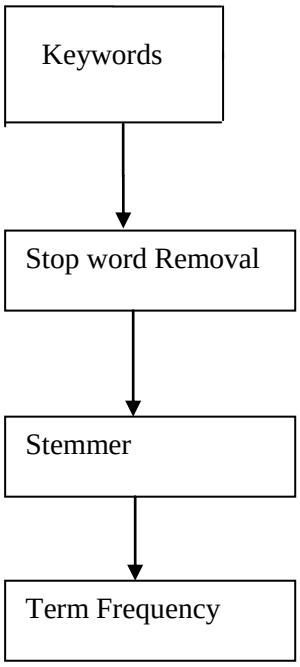

Figure 2: Term based classification

As per schetwz et al.(2008) there exists different types of classifications approaches and each method is distinct from one another. Term frequency and document frequency (tf-idf) is one of the common approach used for documents classifications. Allen et al. (2009) has used the concepts of decision tree for classifying the web page contents. When decision tree is used for classification the results is interpreted as logical relation for viewers understandings. The major drawback of this method is that there is

no enhancement possibility to develop the results of the classification. But, when the documents are classified to a specific category a solution for this existing issue can be addressed. Yan lee et al. (2010) proposed that a user can manually subscribe to a subset of a large number of pre-defined text (news) documents. The set of pre-defined categories is usually static and it corresponds to the categories assigned to the news providing pages when they are first created. In other words, the subscription-based personalization approach is rather straightforward and does not require much classification efforts. Most of the web sites achieve news personalization by adopting the subscription approach, e.g. Newscan-online.

2.3 Recommender Systems

A good number of news recommender systems are available that act based on content and semantics. News Dude [9] is a recommender system that combines both TF-IDF and Nearest Neighbor algorithm. The system considers entire text for recommendation process in case of both short term interests and long term interests. In case of Daily Learner [10, 11, 12] users specify their items of interest. The vector representation of news article is processed with TF-IDF. Using cosine similarity the article is matched with user profile and Nearest Neighbor algorithm is used to analyze the most recently rated news for short term interests [13, 14, 15, 16]. Naïve Bayes Classifier is modeled for long term based interests. Personalized Recommender System (PRES) is based on content based filtering, combining TF-IDF and cosine similarity. User interests are updated whenever the user browses a new item [17, 18, 21, 22].

3 EDUCATIONAL PORTALS

There are numerous education portals available. Each of such portals intends to provide education related news articles. *www.openequalfree.org, www.citylimits.org, www.self.org, www.ngopost.org, www.reapchild.org* etc. are few most prominent education portals. Based on these web portals we have considered a corpus of 4876 web documents. All these documents are pertaining to various categories. The categories we have considered are academic events (workshop, conference, seminar), job fair (online test, interview, evaluation), reading material (journal, video, audio) and admissions (institutions, courses, specialization)

4USER PROFILE CONSTRUCTION

Constructing the profile plays a key role in identifying the interest of a user so that documents can be recommended more accurately. There are two methods of user profile construction namely Explicit and Implicit. In the explicit method, the user is asked to select the interest keywords of his/her

choice from a list of keywords provided. This enables the system to recommend news based on the interest terms. In the implicit method, the browsing pattern is used for extracting user interests.

4.1 Explicit Method

User profile based on explicit method considers users and their interests terms asked explicitly. Table 1 illustrates the details of sample sets of users with their interests.

Table 1: User profile creation based on explicit method

User Number	Interest Terms
U1	I1, I3, I5
U2	I2, I4
U3	I3, I5, I1
U4	I2, I4
U5	I7, I10
U6	I8, I9, I11
U7	I12

In table 1, "I" refers to the interest terms where I1-workshop; I2-conference; I3- seminar; I4- online test; I5- interview; I6- evaluation; I7-journal; I8- audio; I9- video; I10- institutions; I11- courses; I12-specialization. Figure 3 illustrates the recommendation process based on various approaches.

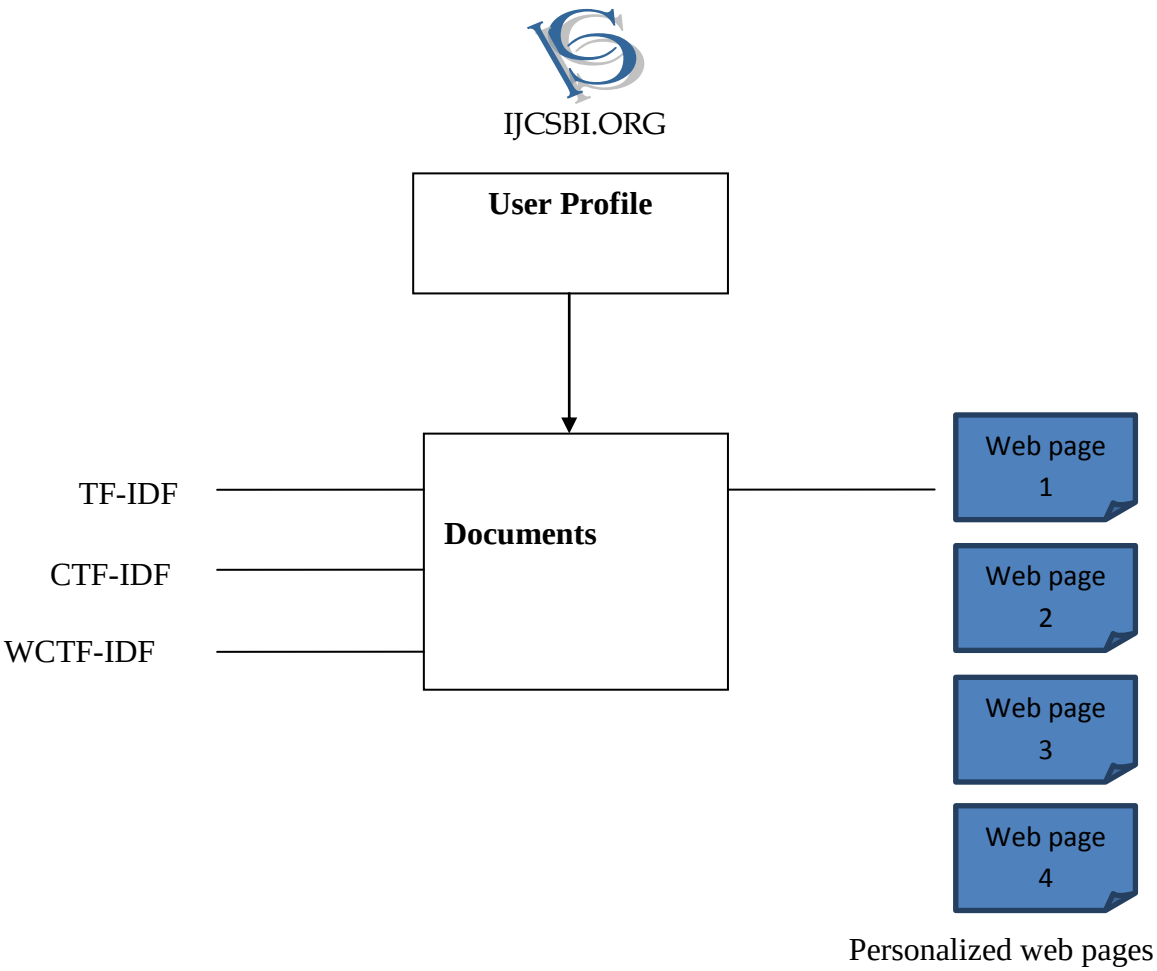

Figure 3. Recommendation by various approaches

4.2 Implicit Method

Based on the browsing pattern the interest of a user is identified. In implicit method of user profile identification, for every item of interest is the corresponding category is taken into consideration.

Table 2: User profile construction by implicit method

User Number	Interest terms
U1	C1i1, C2i3
U2	C1i3, C2i1, C2i2, C3i2
U3	C1i1, C2i1, C3i2
U4	C1i2, C3i6,C4i10
U5	C1i4, C2i12

Table 2 illustrates user profile based interests with their categories. The interest terms are generated based on users long term and short term interests.

Table 3: Categories and their interest terms

Category	Interest terms
C1 (Academic events)	i1-workshop; i2-conference; i3-seminar
C2 (Job fair)	i1-online test; i2-interview; i3-evaluation
C3 (Reading material)	i1-journal; i2- audio; i3-video
C4	i1-nstitutions; i2-courses; i3-specialization

Table 3 illustrates the categories and possible interest terms in the category. In Implicit method for User 1, documents belonging to his/her interest terms are delivered first followed by the other documents belonging to the rest of the interest terms. Likewise for all users the documents are delivered.

5. PROPOSED CLASSIFICATION METHODS

5.1Classification method based on Category Terms (CTF-IDF)

The CTF-IDF recommender primarily uses a vector for each item, and calculates weights for each category terms, instead of going through all the terms. Then, it stores the calculated weights (together with the corresponding terms) of a news item in a vector.

The user profile is also a vector of CF-IDF weights, which can be compared with a news item vector by using cosine similarity. Weights of CTF-IDF are computed as shown below. First we calculate the Category Frequency, cfi; j , which is the occurrence of a category ci in document dj , $n_{i;j}$, divided by the total number of occurrences of all category in the document.

$$cf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \longrightarrow 3$$

Subsequently, we calculate the Inverse Document Frequency.We take the total number of documents, jDj, and divide it bythe number of documents in which the category ci appears,then taking the logarithm of this division, i.e.,

$$idf_i = log \frac{|D|}{|\{d:c_i \in d\}|} \longrightarrow 4$$

Finally, cf is multiplied with idf, forming the weight forcategory ci of document dj.

Therefore,

$$cf - idf_{i,j} = cf_{i,j} * idf_i \longrightarrow 5$$

This change causes the recommender to deal with categoryterms, making it more effective. Another advantageof this method is that the CTF-IDF recommender can processthe news items much faster than the TF-IDF recommender.

5.2 Classification Method based on Weighted Category Terms (WCTF-IDF)

By considering the time spent and frequency of viewing the terms a weighted category term based user profile construction is done. This method refers to a technique that gives a high rank if the interest term appears frequently in a document. In the proposed WCTF-IDF we have considered the distribution of feature terms over various news documents related to education news category. This approach also considers the interest terms that occurs frequently in the documents. For example the interest list (C1I1, C2I1, C1I3) of user 2 means that C1I1 has a high weightage based on time spent on interest term 2 (I2) of category 1 (C1). Subsequently, the other users have equal or lower weights.

The WCTF-IDF formulas is given as

$$WC(t,i) = WCTF(t,i) * IDF(t,i) \longrightarrow 1$$

$$IDF(t) = log[a] * log[b] * log[c]^2 \longrightarrow 2$$

Using equation 2, the classification of a word or term is carried out using the proposed WCTF-IDF approach. The first term namely log [a] is used to calculate a term by identifying the occurrences of that particular word in a document. In the first term the denominator value is low if the term appears more number of times in the document considered for classification.

6. ATHENA

Athena is the extension of Hermes framework which generates recommendations. In order to make effective recommendation Athena framework monitors the behavior pattern of the users. Athena uses several recommender systems especially traditional term based recommender systems in order to compare news item with the profile created. News items are recommended to the users when there exists higher similarity measures.

6.1 Hermes Genesis News Portal Platform (HGNPP)

Hermes framework is an extension to Athena. In order to deliver personalized news documents, Hermes framework is used. In order to retrieve news items from the data source, sematic based approach is followed for making recommendations. A category term is assigned to each news documents so that each news items are considered as input. This input news items are processed internally in Hermes Genesis News Portal Platform (HGNPP) as shown in figure 4 for making personalized news service based on the selection of concepts by the user.

Figure 4: News Articles in HGNPP

6.2 Implementation in Athena

Extension to Hermes framework is Athena, which is used as plug in for Hermes News Portal. In Athena framework there are three tabs which is used for user interface. The tabs are used to browse the news items, recommend the news documents and finally to evaluate the news items that are recommended to the users. The browser in the Athena is used to browse the news items by the user. After reading the news items, user can specify which news items the user is interested. For recommendation the user can click the recommendation tab in Athena. A user is allowed to select only one recommendation tab in Athena. After completion of user activity the Athena analyses the browsing pattern of the user. In Athena, recommendation is also performed based on concept terms. All categories are listed in a category list which is created by each user. Based on the category list, recommendation of news items is done in comparison with the user profile construction.

7. RESULTSANDDISCUSSIONS

Performance comparison is done to identify whether recommendation based on traditional term based approach or recommendations based on our approaches yield better results.Table 4 shows the test results of TF-IDF, CTF-IDF and WCTF-IDF. The averages indicate that CF-IDF seems to perform better than other recommenders on various performance measures. Table 5 shows the test results between TF-IDF and Athena. The difference between CF-IDF and WCTF-IDF regarding recall and the precision is exceptionally large. CF-IDF has good recall value which means that it classifies news items better than other recommenders. Based on these results we have concluded that the recommender system based on category terms performs better precision, recall and accuracy. Performance comparisons of the recommenders are shown in figure 5 and 6.

Table 4. Test Results for TF-IDF, CTF-IDF and WCTF-IDF recommenders

Performance Measure	Traditional Method (TF-IDF)	Category Terms (CTF-IDF) Method	Weighted Category Term (WCTF-IDF) Method
Precision	0.45%	0.92 %	0.79%
Recall	0.19%	0.67 %	0.24%
Accuracy	0.99%	1.87%	1.09%

Table 5. Test Results for TF-IDF and Athena

Performance Measure	Traditional Method (TF-IDF)	Athena Recommendation
Precision	0.79%	0.67 %
Recall	0.67%	0.56 %
Accuracy	1.09%	0.87%

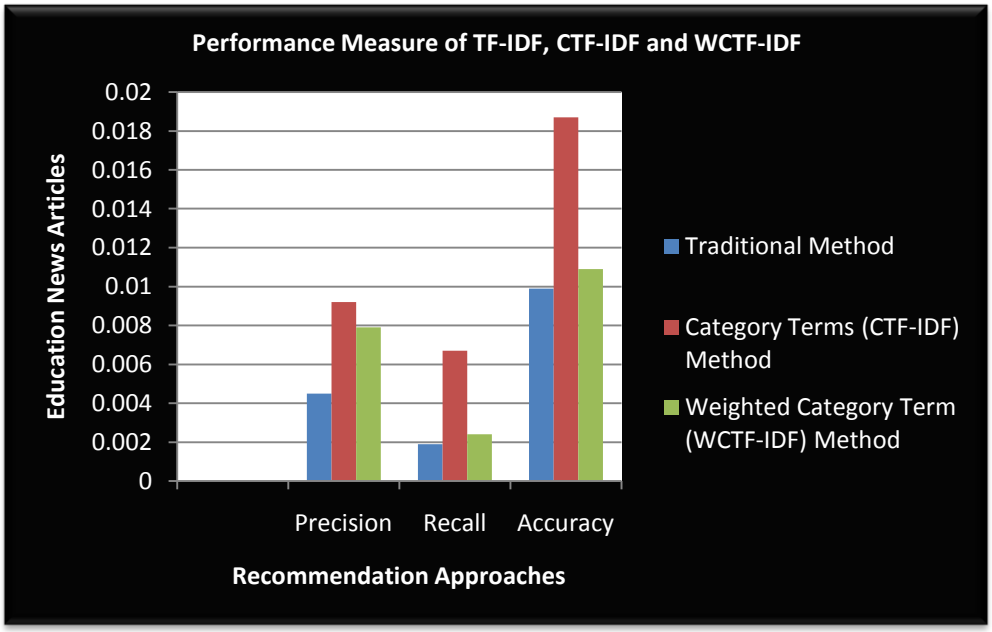

Figure 5. Performance comparisons of TF-IDF, CTF-IDF and WCTF-IDF Recommenders

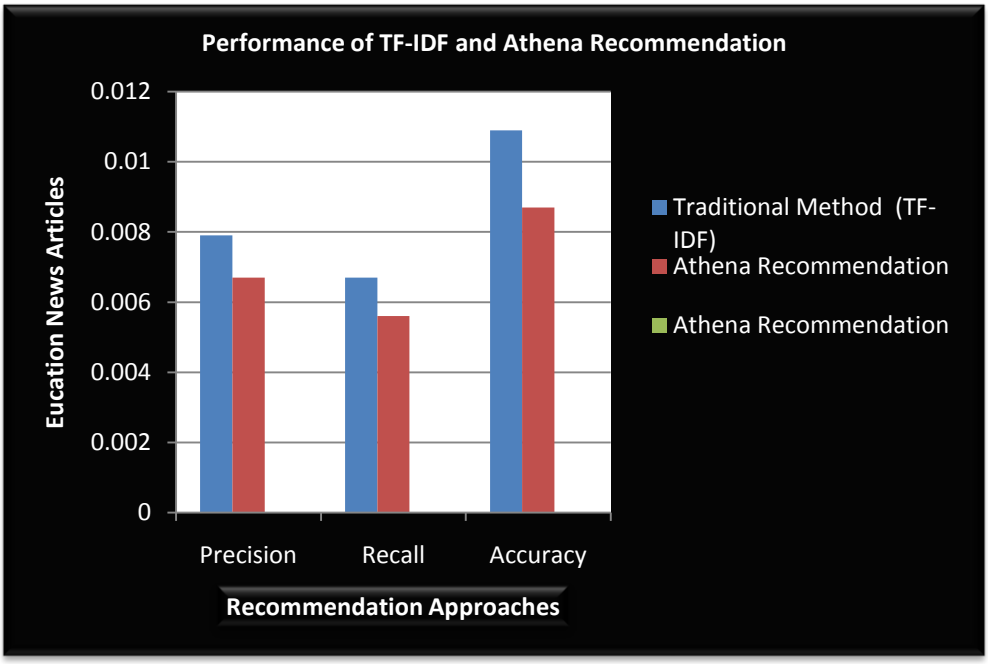

Figure 6. Performance comparisons of TF-IDF and Athena Recommendation

8. CONCLUSIONS

This paper focuses on the improvement of TF-IDF recommendation approach by using the interest terms, categories and weighted category. By employing new classification methods, a better recommendation in terms of recall, precision and accuracy is achieved in CTF-IDF approach. Experimental results show, the CTF-IDF recommender outperforms the WTF-IDF approach and other recommenders on several measures. The CTF-IDF recommender scores significantly higher compared the WTF-IDF recommender on accuracy, recall, and precision values. Based on the results we conclude that there are benefits of using semantic techniques for a recommendation system.

REFERENCES

[1] Chen, C. C., Chen, M. C., Sun, Y, " PVA: A self-adaptive personal view agent system", Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, 2008.

[2] Shawn R. Wolfe and Yi Zhang, "Interaction and Personalization of Criteria in recommender Systems", LNCS 6075, pp. 183–194, Springer-Verlag Berlin Heidelberg (2010).

[3] Jiahui Liu et. al. 2010, "Personalized News Recommendation Based on Click Behaviour", In the proceedings of ACM- IUI'10, February 7–10, 2010, China.

[4] Deng-Yiv Chiu, Chi-Chung Lee and Ya-Chen Pan, "A classification approach of news web pages from multi-media sources at Chinese entry website-Taiwan Yahoo! as an example", IEEE proceedings of the Fourth International Conference on Innovative Computing, Information and Control, pp 1156-1159, 2009.

[5] Carreira, R., Crato, J. M., Gon?alves, D., Jorge, J. A., "Evaluating adaptive user profiles for news classification, Proceedings of the 9th international conference on Intelligent user interfaces, 2004.

[6] Chen, Y-S., Shahabi, C.: Automatically improving the accuracy of user profiles with genetic algorithm. In: Proceedings of IASTED International Conference on Artificial Intelligence and Soft Computing, 2001.

[7] Katakis, I., Tsoumakas, G., Banos, E., Bassiliades, N., Vlahavas, I., "An adaptive personalized news dissemination system", In Journal of Intelligent Information Systems, Volume 32 , Issue 2. 2009.

[8] Chee-Hong Chan et. al. 2010, "Automated Online News Classification with Personalization", In the proceddings of WWW-2009. Italy

[9] Dipa Dixit and JayantGadge, "Automatic Recommendation for Online Users Using Web Usage Mining", In the proceedings of International Journal of Managing Information Technology (IJMIT) Vol.2, No.3, August 2010.

[10] Bardul M. Sarwar, George Karypis, Joseph A. Konstan, andJohn T. Riedl, "Analysis of recommendation algorithms for ecommerce,"in Electronic Commerce, 2000.

[11] R. V. Meteren, M. V. Someren. "Using Content-Based Filtering for Recommendation", MLnet / ECML2000 Workshop, Spain, 2000.

[12] H. Suo, Y. Liu, and S. Cao, "A keyword selection method based on lexical chains," Journal of Chinese Information Processing, 20(6): 25–30, 2006.

[13] ToineBogers and Antal van den Bosch, "Comparing and Evaluating Information Retrieval Algorithms for News Recommendation", InACM Conference on Recommender Systems 2007 (RecSys 2007), pages141–144. ACM, (2007).

[14] Linyuan Yan and Chunping Li, "A Novel Semantic-based Text Representation Method for Improving Text Clustering", In3rd Indian International Conference on Artificial Intelligence (IICAI 2007), pages 1738–1750, (2007).

[15] Flavius Frasincar, Jethro Borsje, and Leonard Levering, "A Semantic Web-Based Approach for Building Personalized News Services", International Journal of E-Business Research (IJEBR), 5(3):35–53, (2009).

[16] Tsoumakas, G., Katakis, I., Vlahavas, I, "Effective and Efficient Multilabel Classification in Domains with Large Number of Labels", In: Proceedings ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08), Antwerp, Belgium (2008).

[17] S. Akhilan and S. R. Balasundaram, "News Personalization Using Enhanced Term – Document Frequency (ETF-IDF) Classification Method" In the proceedings of International Conference and Workshop on Emerging Trends and Technology (2011).

[18] Noy, N. F., McGuinness, D. L., "Ontology Development 101: A Guide to Creating Your First Ontology", Knowledge Systems, AI Laboratory, Stanford University, No. KSL-01-05 (2001).

[19] Sure, Y., Angele J., Staab, S, "OntoEdit: Guiding Ontology Development by Methodology and Inferencing", In: Proceedings of the Confederated International Conferences on the Move to Meaningful Internet Systems CoopIS DOA and ODBASE 2002, Lecture Notes in Computer Science, Vol. 2519. Springer-Verlag, 1205-1222 (2002).

[20] Antonellis, I., Bouras, C. and Poulopoulos, V., "Personalized news categorization through scalable text classification", 8th Asia Pacific Web Conference (APWEB '06), (2005).

[21] S.Akhilan and S.R.Balasundaram, "Enhanced Term Document Frequency Classification Approach for Personalizing News Items", International Journal of Computer Applications, number 2-article 1, published by Foundation of Computer Science, March 2011.

[22] S. E. Middleton, N. R. Shadbolt, and D. C. D. Roure, "Ontological User Profiling in Recommender Systems", ACM Transactions on Information Systems,22(1):54–88, 2004.