

Personalized News Recommender System

Vaibhav Chauhan <i>Student</i> Atharva College of Engineering Mumbai University Malad, Mumbai, India vamb.c95@gmail.com	Yash Dodia <i>Student</i> Atharva College of Engineering Mumbai University Malad, Mumbai, India yashdodia1246@gmail .com	Kiran Shewale <i>Student</i> Atharva College of Engineering Mumbai University Malad, Mumbai, India k.shewale2510@gmail .com	Aniket Darveshi <i>Student</i> Atharva College of Engineering Mumbai University Malad, Mumbai, India aniket.darveshi@gmail .com	Komal Mahajan <i>Student</i> Atharva College of Engineering Mumbai University Malad, Mumbai, India komalsmahajan@gmai l.com
---	---	--	--	--

Abstract— Reading news online has become prominent as the web services provide access to news articles from various sources across the globe. This project delineates a system that collects news from an eclectic mix of electronic news distributors. This system aggregates news from various HTML and RSS (Rich Site Summary) Web pages by utilizing source specific data extraction programs and tabulates them as per the predefined news groups to ameliorate personalized views through a Web based user interface. Reading news has drastically evolved with the development of the World Wide Web (www), from the conventional way of reading a physical newspaper to access to numerous electronic news sources over internet. Web users are undergoing a drastic advancement and they are making a considerable contribution in this field through reviews, comments, ratings, sharing, tagging, etc. Our Personalized News Recommender System provides the users with access to the news articles gleaned from different Web sources. It takes the content from miscellaneous and heterogeneous online sources such as online news portals, bloggers, websites, etc. News items are categorized distinctly according to the user's tastes and the KEYWORD input given on the various categories of News articles. Recommendation and filtering of online news has received much attention in web and artificial intelligence, giving a concise summary of online information for users.

Keywords— *News Recommendation; RSS Feeds; Keyword Searching; Personalization.*

I. INTRODUCTION

A lot of people like to view and analyze news from various news sources. At instances people are looking only for the for the pinnacle news tales in their categories of interest. Hence are more likely to subscribe to RSS feeds from different news sources. Therefore, the users have to scan through all the top news stories in order to read stories of their likes and interests. Just like, a person interested in sports related top news category has to go through all the top news stories analyzing news from different channels. We identified this need of bringing together news from different sources and categorizing them and presenting them to the users as a single news feed. Many users tend to subscribe to RSS feeds of their interest in order to get updated with the latest news. However, many a times this information is scattered across various news sources and spans more than one domain. Our system provides RSS feeds that presents all the news items from various news sources and groups them into categories with the main objectives such as:

[1] Providing the User with personalized news they like by analyzing the User's click behavior.

[2] Processing RSS feeds (representing news channels) and obtaining a single, well-categorized output feed.

[3] Analyzing and searching keywords entered by user to find a particular news article.

To make an application which responds quickly to user's action and preference and provide them better results. Its major concern is to save time and find and categorize top news as per user's interest. Our project aims at processing multiple RSS feeds (four news channels) and obtaining a single, well-categorized output feed.

II. LITERATURE SURVEY

This section gives an overview of existing technologies, their methodologies. We present a comparative study on different approaches.

People usually want to collect more information about a news. Gathering all these News helps users to aware of the current reality. Web blogs are full of un-indexed and unprocessed text that reflects the heterogeneity. It is not easy to walk through a lot of news and read it carefully. Sometime news is directly talked about the product and sometime reviews are explicitly mentioned. Thus, there is a need to collect and process different news sources so that it can be used in decision making processes.

Personalized News Recommender System (PNR) is basically a web application which requires user to log in or register. As the user logs into the application, he/she is required to select the category of news of his/her choice. After selecting the category, all live news of that category are fetched from different electronic news portal through RSS feeds. User can even manually search the news of his/her choice by typing the keyword. On the server-side, administrator uploads the news articles which are stored in the database. Thus, user can even watch the news when there is no internet connection. Hence, Personalized News Recommender System (PNR) is a true replacement to the old recommender system which used only data sets to recommend the news.

There are some topics that work under the umbrella of PNR and have attracted the researchers recently. In this subsection,

few of these topics are presented in some details with related articles.

This work by Dr. M. Durairaj and K. Muthu Kumar was carried out in June 2014. In this work, the first section discusses different News Recommendation System and its functionalities as well as the technologies involved in these systems. The second section discusses about extensively different topic analysis models and technologies involved to develop these models. The third section discusses the advantages and disadvantages of these systems as described by the respective authors. Finally, the paper is concluded with the suggestions and recommendations for building an effective news recommendation system based on the observations made from the extensive reviews and study. The observation made out from this literature study is that the news recommendation is challenging due to the rapid evolution of topics and preferences. The comparison and analysis are carried out on different techniques and methods used for mining the news recommendation models.

Disadvantages- The proposed system does not have good efficiency in terms of user keywords and search results.

This work was done by Florent Garcin, Kai Zhou, Boi Faltings and Vincent Schickel in 2012. This work considers 3 kinds of recommender systems: collaborative filtering at the level of news items, content-based recommendation where we recommend items with similar topics to what was read, and a hybrid where collaborative filtering is applied at the level of topics. Collaborative filtering recommends items to a user based on users with similar tastes, while content-based techniques create recommendations by analyzing the content of the items. Collaborative recommendation compares reading histories in order to extract reading behavior patterns. It recommends news items that other readers with similar reading histories have read. Readers are in different stages at a point of time, and news feeds are generated on basis of transition probability from one stage to another. In conclusion, it demonstrated that personalized recommendations using collaborative filtering can be useful even for individual newspaper sites with limited amounts of data about their users. Disadvantages - The content-based and hybrid recommendations have surprisingly poor performance.

This work has been done by Xindong Wu, Fei Xie, Gongqing Wu and Wei Ding in the year 2011. In order to get real time updates of general Web news topics, interests and preferences of a user, a keyword knowledge base is maintained. The non-news content irrelevant to the news Web page is filtered out. Topics consists of keywords that are extracted using lexical chains which signifies semantic relations between words. Text summarization, collaborative filtering and recommendation of online news have received much attention in web and artificial intelligence, focusing on finding relevant and interesting news and also summarizing concise content. The Personalized News Filtering and Summarization system (PNFS) consists of two phases which are Personalized Web News Filtering and Web News Summarization. The purpose of keyword extraction is two-fold. First, it gives a concise form of the news to the user that saves the reading time. Second, the extracted keywords are also used to build a user interest model. In this work, it has presented the recommendation and summarization components of the personalized news filtering and summarization (PNFS)

system. For the recommendation component, it has designed a content based news recommender that automatically obtains Word-Wide-Web (WWW) that is online news from the Google News portal and recommends news to users as per their preference.

Disadvantages- The Google news in Personalized News Filtering and Summarization system (PNFS) provides general news avoiding missing important news.

III. METHODOLOGY

Personalized News Recommender System (PNR) is a web application which requires users to log in. The user logs into the application and is required to select the category of news of personal choice. Hence after selecting the category, all the live news is fetched from different electronic news portal through RSS feeds relating to that category. Users can manually input search keywords to fetch news of his/her choice. To resolve the problems of existing work we propose a web application that will collect, parse, process, annotate and analyze different news from various RSS feeded channels which express either positive or negative information through crawling.

RSS Feeds: are used to fetch the news from heterogeneous news sites and portals which is stored in the database.

Web Application: The web application is the front end interface for the User to access the system.

End Users: The Users that are accessing and using the system. Users can use the system as soon as they login.

IV. IMPLEMENTATION

The proposed system focuses on the below mentioned steps as also seen in Fig 1 thus showing the flow of mechanism in this system. The proposed system has 6 modules and each of the 6 module has some functionality which eventually leads to the output of the system. Each module is dependent on the other module. The modules are as follows:

A. User Module

In this module, the user runs the Application of Personalized News Recommender. In the application the user is asked to log into the Application or to register to the Application. After the User logs into the application, the user selects the News Category or searches news manually by entering keywords.

B. RSS Feeds Module

In this phase the user selects the category of news of his/her choice. After that, RSS Feeds is used to fetch live news from various online electronic news portal. The RSS Feeds module sends back the results to Application Module where user can see the news recommended by the system.

C. Web Server Module

This module is used to fetch each and every web page content which are stored in the database. User enters the keyword of his/her choice in the search box and web server is used. This module uses word weightage algorithm to fetch the weights of every document/web page in relation to the keyword entered by the user. Word Weightage algorithm uses

TF-IDF technique to fetch the news article from the database. TF-IDF technique calculates the term frequency for every word of the individual web page stored in the database. By using this word weightage algorithm, it selects the document/web page of the highest weight and sends back the results to application module.

D. Algorithm Module

The algorithm used in this system is Word-weightage algorithm which is based on Tf-idf technique. Term frequency-inverse document frequency is abbreviated as Tf-idf, and the tf-idf weight is often used in retrieving data and text mining. Normally, the Tf-idf weight formed by two terms: The first term computes the normalized Term Frequency (TF), which means how many number of times a word appears in a document, divided by the total number of words in that set of words or a document. The next term is Inverse Document Frequency which is abbreviated as IDF, which is computed as the logarithm of the no. of the files or docs in corpus, divided by instances i.e., number of files or documents where the precise term seems. Hence this can be calculated by computing the following: $IDF(term) = \log_e(\text{total no. of documents} / \text{No. of documents with that 'term' in it})$. After calculating Term Frequency (TF) and Inverse Document Frequency (IDF), it selects all the top pages of the results which are having highest weight and sends back the results to the user.

E. Database Module

The Database Module is used to store all the news articles/documents/webpage. All the articles/documents/web pages are stored in the database by the administrator from the server-side application. All the webpages/news articles are fetched from database by the web server and the results shown to the application module.

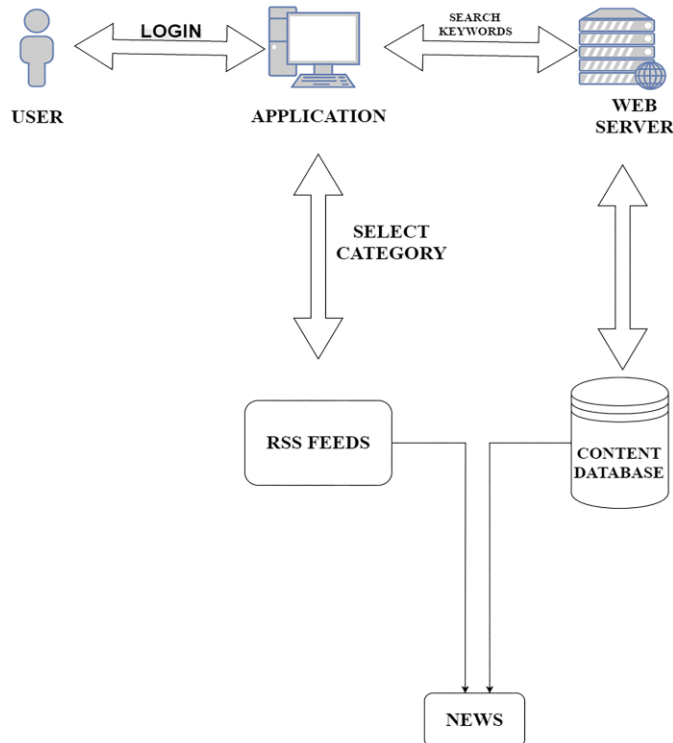


Fig. 1. Proposed System Diagram

F. Application Module

Web application is the basic user interface of Personalized News Recommender System. It is the front end of personalized news recommender system. It shows all the news articles which are recommended through the RSS Feeds and through database. It has search box which is used to search specific keywords other than those which are mentioned in the predefined category of news recommendation.

V. CONCLUSION

Thus we propose to develop a personalized news recommender system based on RSS feeds and Keyword search using word weighted scheme. We ensure that the database updates are carried at regular interval of time. By giving Users what they like from the earliest starting point this application is expected to guarantee that individuals stay updated with what is going on around them. This framework gives a solitary RSS channel that introduces all the news objects from different diverse news sources and classifies them into different categories for the user to experience personalization. Previous recommender systems used to fetch the data from the already stored data sets and it was not necessarily live news. Thus we propose an application which recommends the right news at the right time to the user.

REFERENCES

- [1] Dr. M. Durairaj, K. Muthu Kumar "News Recommendation Systems Using Web Mining: A Study" International Journal of Engineering Trends and Technology (IJETT) Volume 12 Number 6, June 2014.
- [2] Yuqi Wang, Wenqian Shang "Personalized News Recommendation Based on Consumer's Click Behaviour" 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Vol. 1, pp.634-638, 2015.
- [3] A. Balahur, R. Steinberger, M. A. Kabadjov, V. Zavarella, E. van der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva "Sentiment Analysis in the News" Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC), pp. 2216-2220, May 2010.
- [4] Kyo-Joong Oh, Won-Jo Lee, Chae-Gyun Lim, HO-Jin Choi "Personalized News Recommendation Classified keywords to capture User Preference". 12th International Conference on Advanced Communication Technology (ICACT), 2014.
- [5] J. Liu, P. Dolan, and E. Pedersen, "Personalized news recommendation based on click behavior" 15th Int. Conf. on IUI, pp. 31-40, 2010.
- [6] Xindong Wu, Fei Xie, Gongqing Wu, Wei Ding "Personalized News Filtering and Summarization on the Web" 23rd IEEE International Conference on Tools. 2011.
- [7] Florent Garcin, Kai Zhou, Boi Faltings, Vincent Schickel "Personalized News Recommendation Based on Collaborative Filtering" IEEE/WIC/ACM International Conferences on Web Intelligence, 2012.
- [8] S. G. Esparza, M. P. O'Mahony, and B. Smyth "On the Realtime Web as a Source of Recommendation Knowledge", in *RecSys 2010*, September 26-30 2010. Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].