Article

Addressing the Limitations of News Recommendation System: Incorporating User Demographic for Enhanced Personalization

Zerihun Olana Asefa * and Admas Abtew

Posted Date: 18 July 2023

doi: 10.20944/preprints202307.1001.v3

Keywords: News Recommendation system, Cold start Problem, Hybrid Approach, Demographic information, New Users, Popular News

Article

Addressing the Limitations of News Recommendation Systems: Incorporating User Demographics for Enhanced Personalization

Zerihun Olana Asefa * and Admas Abtew

Jimma University, Department of Information Technology
* Correspondence: zrealworld43@gmail.com

Abstract: News recommendation schemes utilize features of the news itself and information about users to suggest and recommend relevant news items to the users towards the interest they have. However, the effectiveness of the existing news recommendation scheme is limited in the occurrence of new user cold start problems. Therefore, we designed a news recommender system using hybrid approaches to address new user cold start problems to ease and suggest more related news articles for new users. To achieve the objective mentioned above, user demographic data with a hybrid recommendation system that contains the scheme of both content-based and collaborative filtering approaches is proposed. To evaluate the effectiveness of the proposed model, an extensive experiment is conducted using a dataset of news articles with user rating value and user demographic data. The performance of the proposed model is done by two ways of experiment. So, the performance of the proposed model performs around 68.05% of Precision, 42.46% of Recall and 52.1% of the average F1 score for the experiment based on individual user similarity in the system. And also performs around 93.75% of precision, 40.25% of recall and 56.31% F1-score for the similarity of users based on the similarity of users within the same category which is better than the first experiment.

Keywords: news recommendation system; cold start problem; hybrid approach; demographic information; new users; popular news

1. Introduction

Today, the information available on the Internet is large and uncontrolled for the masses of users. Users need only the necessary information for their purposes from Internet documents. For doing this it needs more specialized systems rather than traditional Information Retrieval systems to provide only necessary and interesting information to users. The best solution proposed for this problem is the recommendation system (RS). A recommendation system is a system which depends on the history of users and the information or documents accessed by users to make future recommendations. Users who want news information should only get information available from news sources [1–3]. The recommendation system uses basically the contents of the items and the user's profile to provide the information. Therefore, the items must be known to the system, and user information is quite important in predicting user interest. Then the system will compute the relation of the items to be provided by the user to the user's history. And, it is possible to provide similar items for similar groups of users by predicting the group's similarity depending on the user's profile history.

Many works have been done previously focusing on other similar areas like recommending popular news, fresh news, and topic-based news. Still, there is a problem in providing relevant information to new readers or to new users who do not have a history of data in the news recommendation system. The cold start problem of new users is a common problem in all recommendation systems due to the lack of information about new users in the system [4]. The lack of available user data in the situation where new users/ users join the system limits the effectiveness

doi:10.20944/preprints202307.1001.v3

of the news recommender system to suggest relevant news to this kind of user. Due to this insufficient user data, there is an occurrence of a new user cold-start problem. This user data is divided into two, explicit data which is collected by the reader's direct activity (i.e., when users give feedback directly to the news articles while she/ he is reading) and implicit data which the system collects by following the reader's activities (i.e., using the reader's link navigation behaviour including the time spent to read articles) in the system. These user data are not sufficient in order to suggest and recommend relevant news articles to the readers in the scenario where a new user joins the system [1,3,5]. Therefore, we proposed to use demographic data of the user for the new user to start the system.

The aim of this work is to recommend news for any reader even for a user who has no history in the recommendation system. The work is used to suggest news to online readers according to their interest of news from available online newspapers. Thus, rather than searching for all newspapers and news, it is good for the system to recommend only the interesting and related news to readers, depending on the user profile history, which is possible with the News Recommender system (NRS). This system will provide only the relevant information for the readers of news depending on their interest of news categories. For example, some readers need only a specific field of news that is more related to their profession. Whereas, others users want to read a more general field (i.e., sports or entertainment). This news recommendation system is used to help the users on their interest of news and to provide updated news for the users based on their interests.

Our proposed system will use the hybrid of Content-based Filtering (CBF) and Collaborative Filtering (CF) approaches with user demographic data to predict the recommended news for readers. The new user data available that is used for the proposed system are the only demographic data that can be taken when registering for the system. As our proposed system is to fix the problem of new user interest of news, the system will fix it and recommend news for the user according to user groups and their patterns of interest. In addition, it recommends popular news by filtering the articles most read by many users and obtaining the highest rate value. And the user should recommend only the categories that s/he is more interested in. For example, if he/she is interested in entertainment news, he/she should have to get from this category and the same for other news categories. The system returns the news to all new users and existing users, which is category-based (i.e., the news to be recommended/predicted will be the category of news that users want to read).

In this work, we investigated and examined three research questions. The first one is, how to model user demographics with a hybrid approach. The second one is how to figure out this impact on the new recommender scheme towards performance metrics such as Precision, Recall, and F1-Score. The last one is, how to model a new recommender system for a new user.

2. Literature Review and Related Works

2.1. News Recommendation System

News is current information that journalists provide through many media such as word of mouth, printed papers, postal systems, broadcasting, and electronic communication. News (is assumed of being an acronym for the four directions of Earth planet which are North, East, West and South) is an enriching source of conveying information on current events and trends presented to the readers. The acronym of NEWS is describing that the source of news can be from the four directions because it is information from around the world. News is the information most accessed by many readers [6].

News articles are published and reach readers on news websites for online users and on printed paper for printed paper readers. The online newspaper is the online standard of news published on the Internet in series, which contains information about current events and other informative articles on different categories, such as politics, sports, arts and advertising [7]. Today, there are many different readers willing to read different newspapers available on the Internet using different sources of online newspapers. In addition, there are many sources of newspapers that provide news for these users. Many newspapers are specified for single categories (i.e., sports, Arts, Vacancy

Announcement, Business, Politics, etc.). Although it is classified, users want to read only the news provided by different news providers for their interest.

News Recommendation systems have evolved as an answer to the information overload problem prevalent with online news readers, while users look for relevant news information out of a huge premise of news content available in online newspapers. Such systems are used to provide recommendations to the users guiding them towards news articles that match their interest and their choice. For this reason, the News Recommendation system is a specific study area under the recommendation systems, where these systems are used to suggest news to users that match their interests and personal preferences. Since much information at different times is published on the news website, there will be a lot of information on these websites. But the user needs the recent news and the system.

2.1. Challenges in News Recommendation System

To recommend relevant news articles for specific readers, it is not easy work and it needs to solve some common problems with news recommendation system. Because the news information should follow the criteria that news should have to include in its system there are some challenges when recommending relevant news articles for readers. So, considering such challenges and making recommendations is the main work of the news recommendation system. As many researchers presented their works in this area, there are many challenges filtered which face in the news recommendation system. Some of the challenges studied in the survey work of [8–12] are discussed below. Some of the common methods for all kinds of recommendation systems are explained earlier.

1. Scalability: The volume of news is large since it is collected from different sources of online newspapers with different categories of news within a short period of time.
2. Cold start problem: It is a common problem of the recommendation system in many domains of the recommendation system which occurs because of the lack of information about new users and new news articles for the news recommendation engine.
3. Data sparsity of user profiles: Most news readers are not willing to provide their profile for the news recommendation engine for the sake of privacy since the information they need is about news which contains a series of information they read like politics or other serious news.
4. Freshness of news articles: The special characteristic of news is the information readers feed from it is fresh and the current activities are done. But if the fresh news is not relevant to the users the system will not recommend the breaking news.

The challenges listed above are the main ones in the news recommendation system, and many studies have been done to fix them even though they are still not complete. In our study, we propose a solution to fix the cold start problem to overcome the problems of new users. Since new users have no history in the system, most systems will not recommend related news to users. Therefore, readers will read the news they will not like to read.

2.2. Approaches of News Recommendation System

The usual approaches used in the News Recommendation System (NRS) are three. These basic approaches of recommendation systems are Content-based Filtering, Collaborative filtering and Hybrid (which is a combination of the first two approaches). So, the details of these techniques were discussed.

2.2.1. Content-Based Filtering (CBF)

In our study, we have used different techniques as discussed above. From these techniques, one is the content-based filtering approach. Content-Based Filtering (CBF) is one of the techniques of a news recommendation system, which is based on the properties of the items and tries to recommend news items which are similar to those a given user has liked to read or rated in the past. This method finds the preferences of the current user about the articles of news using the rating history of the current user related to previously used items.

The similarity of items is determined by measuring the similarity in their properties. So, in this type of filtering method, there is no dependency on the rating records of other users in order to generate preferences for current users. Content-based systems focus on the properties of items. For example, if the user has purchased a book on Amazon.com which uses a recommendation system, then the user starts getting additional preferences for buying books from online book store which includes the same or similar keywords information for books.

Therefore, CBF-based recommendation systems generate recommendations using comparative representations of content relating an item to representations of content that are interesting to the user. In this method, news recommendation is done based on the content similarity between the news articles that the user has read, with the newly published news by considering news recent times. As such, many methods have classified this issue as an information retrieval (IR) issue, where content related to the user's preferred choices is considered an enquiry, and unrated documents are evaluated on the basis of relevance/similarity to this enquiry [13–15]. In order to recommend one particular news to a user, the content-based approaches will get the previously rated news based on their category for example, (sports, science & technology, health, business) and then the news with the highest similarity to user preferences are recommended.

2.2.2. Collaborative Filtering (CF)

The other technique used in our study is collaborative filtering. In Collaborative filtering-based RS, the user will be suggested items that people with similar interests and preferences liked in the past. In a CF recommendation system, in order to suggest items to the user, the collaborative filtering recommendation system looks for the peers of the user, that is, a set of users that have similar interests in the item. Then, only the items that are the most liked by the peers of the user would be suggested [4]. So, collaborative filtering is a technology that aims to learn user preferences and make recommendations based on user and community data. It is a complementary technology to content-based filtering (e.g., keyword-based searching). Probably the most well-known use of collaborative filtering has been by Amazon.com for example where a user's past shopping history is used to make recommendations for new products. However, various approaches to collaborative filtering have been proposed in the past in the research community.

In addition, Collaborative filtering (CF), as a kind of personalized recommendation technique, has been widely used in many domains [16,17]. However, collaborative filtering also suffers from a few issues, for instance, cold start problems, data sparsity, scalability and so on. These problems seriously reduce the user experience. Collaborative filtering recommends items to users according to their preferences. Therefore, a database of user history must be available. However, the database is always very sparse; that is, the user only rates a small number of items. Up until now, there have been many researchers who have focused on prediction accuracy and proposed some solutions.

CF has been becoming popular in both the academy and industry fields with great speed. Despite the overall success of CF systems, they suffer one serious limitation, namely the cold-start problem [5]. This cold-start problem includes two major aspects: new user and new item. Before a recommender system can present a user with reliable recommendations, it should know about this user's preferences/interests, most likely from a sufficient number of behaviour records, e.g., ratings or log archives [18]. Since collaborative Filtering uses the similarity of users with respect to the items, they rated it is based on both user and item. Collaborative filtering is a technique that uses user-based similarity or item-based similarity. User-based Collaborative Filtering (UBCF) is one of the widely used recommendation technologies, remaining due to its convincing simplicity and quality of recommendations. It starts with the assumption that if a group of users have similar interests in their past, they will have similar interests in other items in the future [19]. The basic idea of User-Based Collaborative Filtering is to find a group of users who have a history of agreeing with an active user (i.e., they either gave similar ratings). Once a neighbourhood of users is identified, feelings from these neighbours are combined to produce recommendations for the active user. Item-based Collaborative Filtering (IBCF) is the similarity between items that are decided by looking at how other users have rated them. And it only considers users who have rated both items for each user and it calculates the difference in ratings for the two items finally, it takes the average of this difference over the users. For example, news in a similar category rated by different readers should be considered as similar items.

2.2.3. Hybrid Approach

In order to exploit the strengths of content-based and collaborative recommendations, one simple method is to accommodate both methods to obtain the two separately-rated lists of recommendations, and then arrive at a final list, which is a merger of the two results. The two predictions employing an adaptive weighted average are combined, where the weight of the collaborative component increases in line with the increase in the number of users accessing an item [20,21]. In this study, a hybrid approach is used which contains both collaborative filtering and content-based techniques.

The single recommendation system approach is not efficient enough to generate relevant and accurate recommendation preferences. So, hybrid recommendation systems came into existence to overcome the limitations of traditional recommendation approaches. The reason why the hybrid approach was selected is that when these two techniques work together it will increase the performance of the system and enhance the capability to answer user's interest. These systems are based upon combining advantages of more than one traditional approach for recommendation generation for example; collaborative filtering approach with a content-based approach or collaborative filtering with demographic characteristics-based recommendation approach [22]. The Figure 1 shows the Recommendation System Taxonomies.

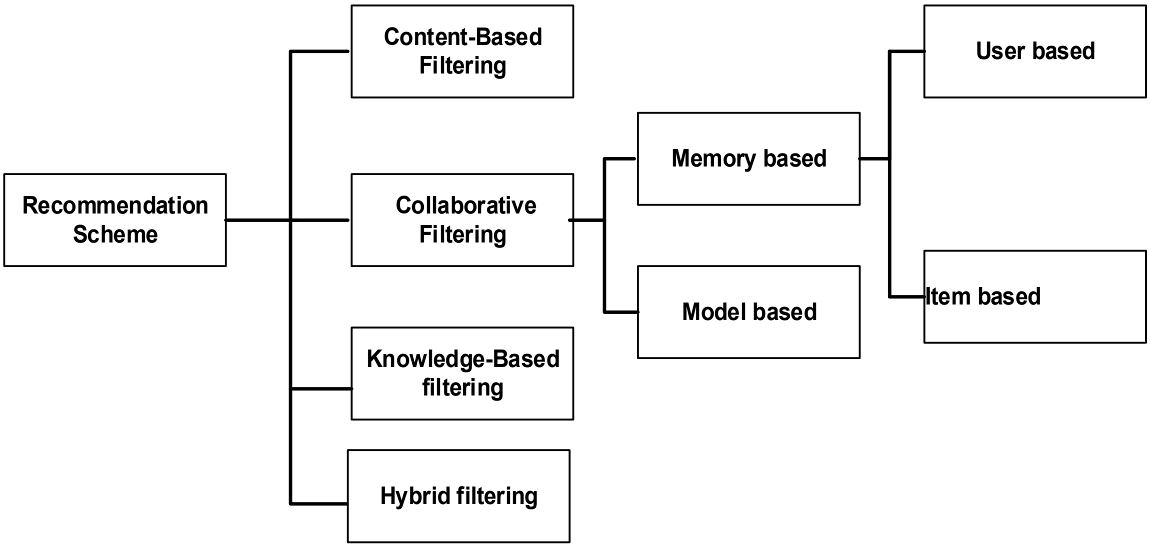


Figure 1. Recommendation System Taxonomies.

2.3. New User Cold Start Problem

There are many challenges of the Recommendation system as discussed in section 2.2, out of them the cold start problem is the common one which happens when user data or item data is scarce. The cold start problem is defined as the problem that happens with the lack of information in the transaction of the recommendation system. The transaction in the recommendation system happens between the interaction of the user with the product or information to be recommended for the users. So, the cold start problem occurs when a lack of new items or new users for the system is announced [23]. Similar to the new user problem, new items which have not been rated could not be recommended, which is referred to as the new item problem. Since a new user, does have not data in a system, normally the system cannot provide satisfactory recommendations.

The main purpose of our study focuses on the new user problem which is a key issue that determines the initial success of users. More accurate recommendations for new users could make these new users stay rather than push them to the competitors' sites. The new user problem in the recommendation system is described as the period from the moment when a user joins the system to the moment when there are enough ratings to yield a stable list of neighbours (i.e., users with similar preferences) [23].

Various researchers on hybrid recommendation systems combine both content information and rating data [24–26] to address the new user problem, where content-based similarity is used for new users or new items. In most currently used systems, demographic information is used as users' attributes to calculate similarity among users. For example, Pazzani [27] uses the gender, age, area code, education and employment information of users in the restaurant recommendation application. So, in our case, we used the similarity of users according to their demographic information.

2.4. Related Work

The performance and effectiveness of hybrid filtering approaches would be improved by combining two or more features. A comprehensive metric helps the scheme to suggest news articles to the readers based on their preferences.

The study conducted by Zhongqi Lu et al. [28], on Content-based Collaborative Filtering for News Topic Recommendation, brings both Content-based Filtering and Collaborative Filtering approaches together. The Content-based Filtering approaches inspect rich contexts of the recommended items, while the Collaborative Filtering approaches predict the interests of long-tail users by collaboratively learning from the interests of related users. In this experiment and analysis, the performance gains and insights into news topic recommendations in Bing were discussed. This work brings both the Content-based Filtering approach and the Collaborative Filtering approach together to make recommendations. Intuitively, the key to both approaches is to find similarities and do clustering implicitly. Content-based Filtering approach relies on the similarity of contexts and clusters the items, while the Collaborative Filtering approach finds similarity in user-item links and clusters the links.

The recent study by Hee-Geun Yoon [29], entitled Personalized News Recommendation using User Location and News Contents focuses on a specific user based on the preference of the user. With the increasing use of hand-held devices, the interests of users are not influenced only by news content, but also by their location. The proposed model is the Spatial Topical Preference Model (STPM). STPM is trained only with the news articles that the user actually reads. Therefore, the final proposed model is a combined model of STPM and LDA. In the evaluation of their model, it is shown that STPM reflects user locations into news article recommendations well, and the combined model outperforms both STPM and LDA. These experimental results prove that location-based user preference improves the performance of news article recommendations, and the proposed model incorporates the locational information of users into news recommendations effectively.

A. Darvishy *et al.* [30], studied New Attributes for Neighbourhood-based Collaborative Filtering in News Recommendations. In this study, the authors utilized new attributes in addition to standard recency and popularity such as Reading Rate and Hotness. These attributes are defined in the user profile and news metadata and are used in neighbourhood-based collaborative filtering in news recommendations. The authors analysed the proposed attributes in the user profile construction and the news metadata enrichment by exploring similar users' interests in news reading. This is carried out via experiments using k-means. Then they compared the precision, recall and F1-score in a series of experiments to evaluate the news recommendation with these attributes. The experimental results show that the proposed attributes improved the accuracy of news recommendations with higher precision and F1-score. They conclude that Reading Rate and Hotness in news have a significant impact on personalized news recommendation systems.

In the work of Sunitha and Adilakshmi [31], they proposed a new approach to using the user's side information in addition to the user-item rating matrix to address the new user cold-start problem. User's side information is obtained from Social Networks. First CF method is used to form user clusters based on the similarity of users. User's side information is obtained from social networks and the same is used to form a social matrix. Finally, combine user history with a social matrix to provide recommendations to new users. The experimental results proved that the performance of RS is improved over traditional CF systems. In this study, they have proposed and implemented an algorithm which will use both collaborative filtering and social network information into account in order to improve the accuracy of a Recommender system to address the Cold-Start problem. Due to the exponential growth of the Internet, users are facing the problem of Information overloading. Recommender Systems (RS) serve as an indispensable tool to solve information overloading problems.

In general, using the features of readers and news contents, based on the news topic the recommendation can be provided if the news have similarity based on their topics. In addition, the user's location is used to provide recommendations since users in similar locations should have similar interests in the news. And also, the news time is used to provide recommendations by considering the hot news that should be read by any readers. But all this work didn't consider the problem of new readers who have no historical data in the recommendation system. So, we consider the new reader problem as our main problem and fix it in this study.

The Hybrid approach with demographic data is proposed for news recommendations of the new user to recommend the new user who has no historical data in the system. This hybrid approach with demographic data uses the combination of both content-based filtering and collaborative filtering techniques of recommendation with user demographic data. The demographic data submitted to the system by the user is used to make effective the recommendation provided for new users with no history since the new user has no history in the system. This proposed model attempts to solve the problem with new users by categorizing the new user with the existing user based on the demographic data they register to the system.

3. Methodology

Our work proposes a hybrid approach including the features of both Content-Based Filtering and Collaborative Filtering approaches with user demographic information and popular news recommendation. The hybrid approach in our work is to improve the shortcomings of both approaches by considering popular news to provide actionable recommendations to news readers by solving new user cold start problems.

3.1. Data Set Preparation

We used the user demographic data, news articles and rating data of news articles by users. The data set used for the evaluation of the proposed system is collected from the news dataset which is available on the Internet and published in the English language from popular GitHub. This website collects data from Seven Search Engine websites, and the data collected from the website consists of information on the features of news articles. Another data set we used is the demographic data set prepared by ourselves and the rating data set collected by interacting with users and news articles.

(a) **User demographics data set**

It is data of users which includes user age, user location, user ID, user occupation and sex. It is important in our work since our work is purposely based on the user problem and we have to have their demographic information. The dataset contains four attributes; those are User ID, Age, Occupation and Sex. Generally, user demographic information is the information used for finding similar users and categorize them according to their demographic information similarity for this work.

(b) **News dataset**

The news articles data were collected from the Seven Search Engine Websites with their URL, Categories of news, Dates published, sources, and titles. Let us describe each of their attributes and their function in our system. We have changed the text data of the source news into a table in our database according to their attributes appropriately. The news dataset is the item part of the recommendation system in our news recommendation system to be recommended for the news readers by the recommendation system we proposed.

(c) **Rating value data**

Rating is the value given for some online items by the user of the system for commercial purposes and for reading or watching videos. In our case, the rating data is the data set that holds the rating value given by the users for specific news articles by the existing users for existing news articles in the data set. The table of this dataset in our database consists of the user Id, the rating value which is 1, 2, 3, 4 and 5, the news URL, the Category of the news and the date of the news. They are discussed below.

Pre-Processing Data Set

The data set we used for evaluating our work is huge and we need to pre-process by using an appropriate algorithm used for pre-processing all data sets we have done for each as follows. This reduces and handles the scalability problems in our proposed News Recommendation System.

i. **Categorizing Existing User**

We categorize users to pre-process the existing user dataset and to find similar users based on their demographic data by using the similarity of user demographic of data in the database. We categorized users based on their professions, sex and age. We select these attributes because we have the assumption that people with similar professions have similar interests in the news because people want to get information about their professions every time. In addition to this professional data, we categorize users according to their sex, since males and females have different interests in the news for their personal life. For example, female users want news of fashion for females and many news related to female users. Then, age also has an influence on user interest as young users and old ones have different interests in the news. Finally, we have users categorized into 38 groups based on the 19 professions with each to both female and male sex.

ii. **Categorizing a news dataset**

The second dataset we used in our work is news articles. The news articles data set is many and we need to group the news articles based on news categories to check the similarity articles to be recommended. The news articles in the dataset contain different features like time, categories, titles, and rate values. Therefore, we grouped this news dataset on the basis of the category of news articles. This means that news with similar categories is grouped under the same group. Generally, grouping the news articles reduces the time used to process and the memory used while retrieving the required news articles and unnecessary data. So, this method overcomes the scalability problem one of the common problems in the news recommendation system. Then we have news articles categorized into 7 groups based on the 7 news categories.

iii. **Grouping Rating Data**

The other data set we need to categorize is the rating data of news rated by users. This data set contains the news rated by users with the user ID and the news ID or the news URL. In addition to this, the rating data contain other features like time, rate value, and news categories. Since it contains the data of the user IDs, the rated news and the many news rated by many users, it is big data. Processing these huge data takes many times and memory, which leads to the scalability problem. Therefore, scalability is an important issue we need to consider in the study, we used the grouping data method for this dataset. We grouped this data set based on the rate value of the news given by the users who read and rated the news. After it is grouped into 4 groups based on the rate value ranges, we have 4 groups of news. These are above-average news rating data which contains news rated by users with 4 and 5 rating values, average rating data with 3 rating values, below rating data with 1 and 2 rating values, and visited rating data with no rating value but visited by the readers.

doi:10.20944/preprints202307.1001.v3

3.2. Registering a New User

Registering user enables us to collect new user demographic information. Since our system needs the demographic information of the user to recommend, we submit new demographic data from the user to the system using the user interface we prepared. These collected data of new users are recorded on the user dataset and the users with the most similar demographic attributes are categorized under similar categories. The categorized users also need to be registered for their groups by the system. This data is used for our assumption that the users with the most similar demographic data have a common interest in news articles. So, any new user needs to be registered for the system to have recommendation news.

Categorizing New Users

The existing users were categorized into their appropriate groups based on their similarity according to their occupation, age group, and sex offline before the new user registration process. When a new user registers, the process of searching for the appropriate category of users will continue, and if the category is found, the recommendation will be processed using the history of existing users similar to the registered new users. But if the new user couldn't get the exact group, a new group for the users is created in our system. The registered user groups should be identified according to user similarity and the data are registered to the appropriate category or similar groups. This identification of new users is done based on demographic attributes. If the registered new users' attributes are similar to existing users' demographic data, the users will belong to that categorized existing user. Then the system adds the user demographic data to his/her respective group.

The users categorized under similar users based on their professions and sex may be many users, and we need to reduce the number of users under the selected categories. Filtering this information will reduce the number of users by assuming that users with similar age groups will have a similar interest in news articles to be read. Firstly, we grouped users into four age groups. The first group with an age range of 15-30, the second group with a range of 31-45, the third one with a range of 46-60 and the last group with a range greater than 60. From the four age groups, the new users will be in one group and the news rated by these users in this group is fetched from the rating database.

3.3. Hybrid Recommendation with Demographic Data

As its name suggests, the hybrid technique is the value of different techniques combined together to have better-combined techniques. So, in our case, the values filtered in content-based techniques and the result filtered by collaborative Filtering techniques are combined with the popular news filtered to get the final result. The combination of the result of news recommended based on rating prediction, category-based news and popular news filtered based on the frequency of the articles rated by many readers and the rating values given by the readers. The hybrid of our proposed approach is done by combining the results obtained by predicting and popularity to be recommended after ranking by the time value order.

In the final stages, the top-N articles recommendation is done by the ranking algorithm we used for the news recommendation system. For ranking the articles, we considered the time the news articles were published as the first priority and then the popularity of the articles by the readers. Each of our approaches has used different ranking methods for the articles to be generated. And the results obtained in each individual approach are combined together by aggregating their results. Finally, we generate the news articles to be recommended by the time they were published. This means the recent news is displayed at the top. Finally, the system stores the recommended log files of the system to reduce the search complexity for the next search. Let us explain each recommendation approach and how they are combined by a hybrid approach.

3.3.1. Content-Based Filtering News Recommendation

Content-Based Filtering (CBF) is one of the techniques commonly used in NRS and other RS to make recommendations based on the contents of the items or the information needed to be recommended by using the similarity of the content of the item. So, in the NRS case of our work, news with a similar category is considered as similar news to be recommended for readers of the same news categories. For example, users who read from the sports categories are suggested other news from sports categories. In our work, we consider the news articles rated by users to recommend, and so the users who rated news from a similar category are considered to be recommended the news from that category with the highest rating values.

The news categories are used to categorize the news rated by users and this news is rated by different users. From the user category done based on user demographic data, the news rated by the users in the category should be filtered. Then, from the news filtered, the categories of the filtered news are identified and the category which has the highest number of news articles rated is considered the category that this category of users liked. Then, our system will generate the top news from this category by considering the recency of the news and then the rating values to recommend.

3.3.2. Collaborative Filtering News Recommendation

In this work, the collaborative ffiltering approach considers the assumption of those users who rated or read news articles have the same taste of news for the upcoming news. But, in the case of a new user cold start, the problem occurs with the lack of rating data for the news articles in the system. Therefore, the provided recommendation could not be a personalised recommendation, which means that the recommendation is not based on the interest of the user. So, to find similar users first, we use user demographic data since our problem aims at recommending users for the user who did not have any history on the system.

In most currently used systems, demographic data like user age, profession, location and sex are used as users' attributes to calculate similarity among users. We considered the similarity of the users by their history of the news they rated in their past time which is stored in the rating database. Therefore, new users who do not have a history in the system should be added to the active users. Active users mean, the users who have the rated news and the history information in the system. To add the new users, we need to have the information used to add to the active users based on the similarity. Since we have the assumption that users with similar demographic information will have a similar interest in news articles, we considered the demographic data as the information used to measure the similarity between the active and the new users. So, their similarity is done by categorizing their demographic data.

So, the new user added to the system category is identified, the news rated in this category is filtered and the filtered news is predicted to the new user. But the generation of news for this user should consider the time the news was published and the rating value given by the active users. This approach requires both the online process to automatically category the new user into active users and the offline process to fetch the rated and stored news on the database. Thus, in this proposed work, we applied both memory-based and model-based algorithms. The memory-based is used while our system filters the news from the memory, whereas, the model-based is applied while categorizing the users in online by learning the user's neighbours from the system. So, it is possible to predict the rating values of the new user based on the rating values of the existing users.

Finally, the news rated by the user is filtered into a group according to their demographic similarity. Since the rated news is categorized into four different groups, the system should check all groups and follow the priority to return the news. The news with the highest rating value should receive the first priority if it satisfies the number of news to be generated for the readers. If not, the next group is checked and should also fulfil the determined number of news articles to be retrieved from this group. In addition to this high rating value, we consider the time of news since users need recent news. Therefore, the latest news that has the highest rating value is generated for the readers.

3.3.3. Recommending Popular News

The popular news assumption in this work is the news articles with the highest rating value and rated by many users in recent times. So, we retrieve all the news in recent time days. Then, we check their frequency or the number of occurrences since the articles frequently occurred is the news rated

doi:10.20944/preprints202307.1001.v3

by many users and it is popular with many readers. Finally, we retrieve these articles by checking their rating value and retrieve all with the highest rating value.

3.4. Proposed Model Architecture

The proposed system architecture presents the model and algorithm used to accomplish the task to be performed in the work. Since the problem is to solve the new user recommendation the work includes many components. The entire proposed system architecture is shown in Figure 2.

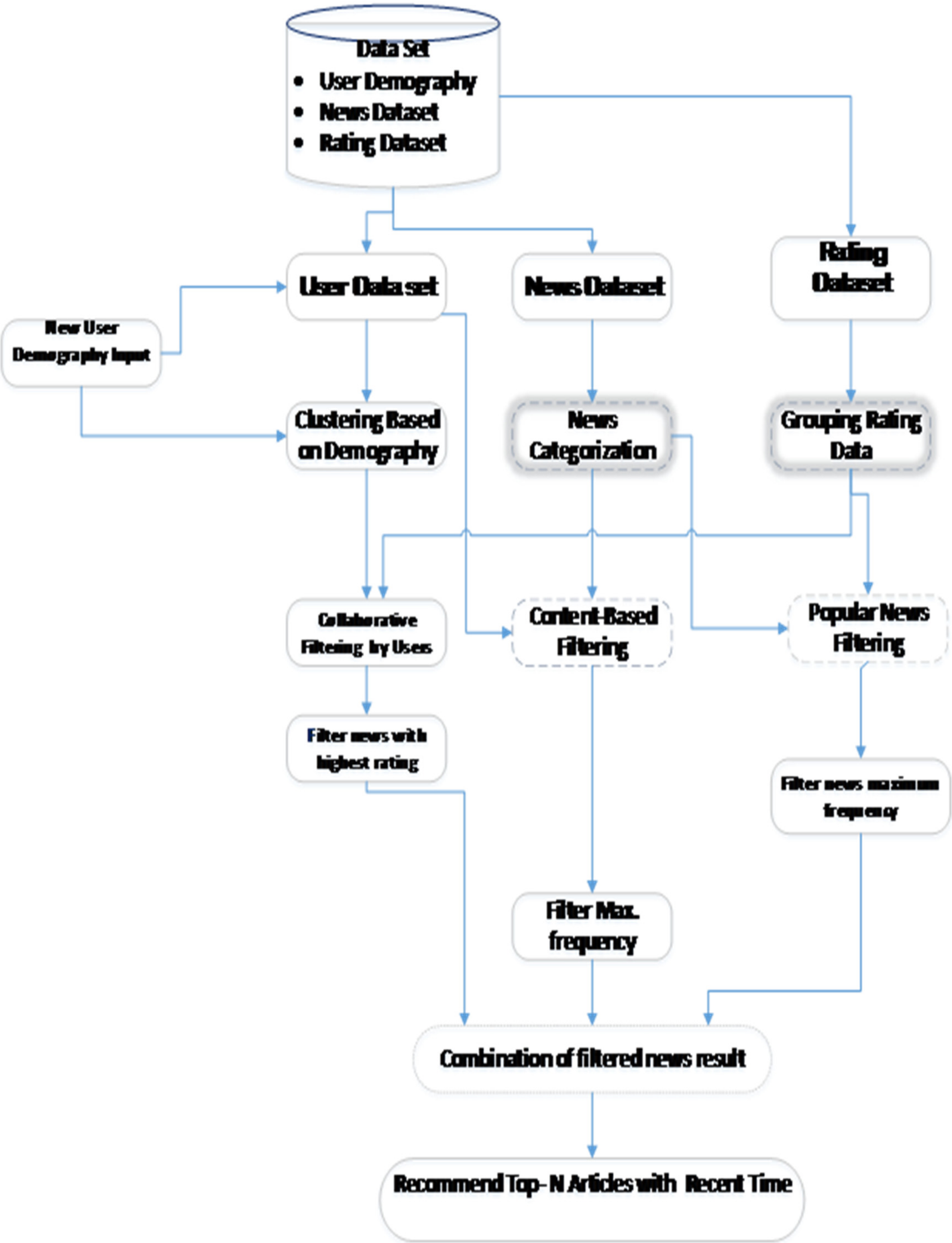

Data Set
User Demography
News Dataset
Rating Dataset
User Data set
News Dataset
Rating Dataset
New User Demography Input
Clustering Based on Demography
News Categorization
Grouping Rating Data
Collaborative Filtering by Uses
Content-Based Filtering
Popular News Filtering
Filter news with highest rating
Filter news maximum frequency
Filter Max. frequency
Combination of filtered news result
Recommend Top-N Articles with Recent Time

Figure 2. Proposed system Architecture.

3.5. Algorithm of the Proposed Model

The algorithm of our proposed system which represent all the approaches we used to overcome the problem formulated in our work is as follows. The pseudo-code consists of each of the separated approaches algorithm and their combination or the hybrid we developed in this study.

Algorithm 3. Pseudo code of Hybrid with Demographic Data Algorithm

```
INPUT: Set of user-news Rating category $R_c$, set of new User Demographic data $U_n$, set of news
category $C_N$ and set of User category $U_C$
OUTPUT: set of news articles
BEGIN
FOR each new user $U_n$
        FOR each of user category $U_C$
        Find new user category $U_{nc}$
        IF (user $U_{nc}$ not exist)//if $U_n$ has no similar user from existing $U_C$
        FOR each news category $C_N$
                Find news with highest frequency $P_{HF}$
                FOR each $P_{HF}$
                        Find news with highest rate value $P_{HR}$
                        Store value on $P_N$
                END FOR
        END FOR
        Display $P_N$ by time descending order
        END IF
        Else
        IF (user $U_n$ age<=45) // user is young
                FOR each News category $C_N$
                        Find news with highest rate value $P_{HF}$
                        FOR each $P_{HF}$
                                Find news with highest frequency $P_{HR}$
                                Store on $P_N$
                        END FOR
                END FOR
                FOR each News category $C_N$
                       Count news articles
                END FOR
                        Find news category with highest frequency $C_{HF}$
                        Store on $CB_N$
        CF News
                        FOR each rating category $R_C$
                                Find rated news $N_R$
                                IF (rated news exist $N_R$)
                                        Store on $CBF_N$
                                END IF
                        END FOR
                Store $P_N$, $CB_N$ and $CBF_N$ on Hybrid News $H_N$
                Display $H_N$ in descending order of time
        END IF
        ELSE // if user is old or greater than 45
                FOR each News category $C_N$
                        Count news articles
                END FOR
                        Find news category with highest frequency $C_{HF}$
                        Store on $CB_N$
        CF News
                        FOR each rating category $R_C$
                                Find rated news $N_R$
                                IF (rated news exist $N_R$)
                                        Store on $CBF_N$
                                END IF
                        END FOR
                        Store $CB_N$ and $CBF_N$ on Hybrid News $H_N$
                        Display $H_N$ in descending order of time
        END ELSE
    END
```

4. Result and Discussion

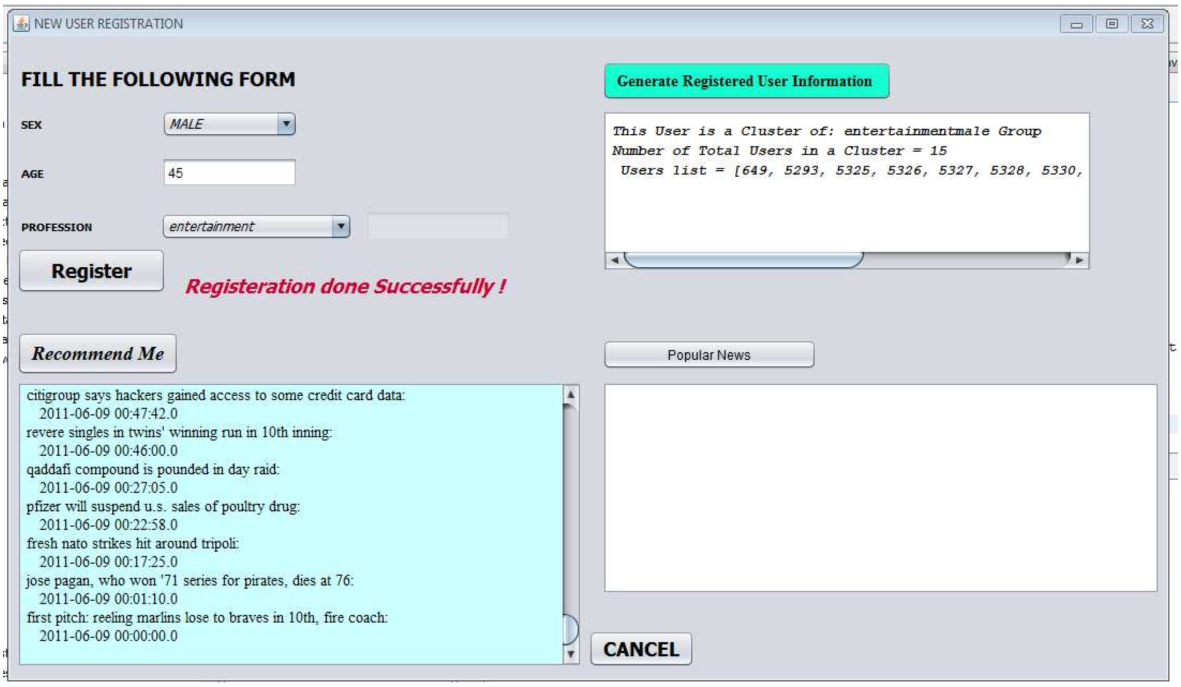
4.1. Results

We used 160 active users which contains different professions, both sex and different age groups. These selected users are the users who rated 32,624 news articles and make 303,594 rating data. Based on this data, the system generates the news for the new user registered and its categories the user into the appropriate category and if the category of the new user is not available in the system it creates a new category and adds the data to it. When a new user registered to the system and asks for a recommendation the system needs the user to request through the user interface prepared for the new user and use the information collected from user's and articles recommendation is done on the interfaces shown in Figure 3.

To evaluate the performance, we depend on the objectives to recommend the more related or interesting news articles for new users we should have to evaluate the relatedness of the news with users. So, the popular and the most used metrics for any information retrieval to measure the relatedness or the interests are precision, recall and F-score. Table 1 shows the descriptions of all metrics with their relationships [32].

Table 1. Evaluation Metrics [32].

	Recommended	Not Recommended
Good Articles	**TP** (True-Positive)	**FN** (False-Negative)
Not Good Articles	**FP** (False-Positive)	**TN** (True-Negative)



FILL THE FOLLOWING FORM

SEX MALE

AGE 45

PROFESSION entertainment

Registeration done Successfully !

```
This User is a Cluster of: entertainmentmale Group
Number of Total Users in a Cluster = 15
 Users list = [649, 5293, 5325, 5326, 5327, 5328, 5330,
```

```
citigroup says hackers gained access to some credit card data:
  2011-06-09 00:47:42.0
revere singles in twins' winning run in 10th inning:
  2011-06-09 00:46:00.0
qaddafi compound is pounded in day raid:
  2011-06-09 00:27:05.0
pfizer will suspend u.s. sales of poultry drug:
  2011-06-09 00:22:58.0
fresh nato strikes hit around tripoli:
  2011-06-09 00:17:25.0
jose pagan, who won '71 series for pirates, dies at 76:
  2011-06-09 00:01:10.0
first pitch: reeling marlins lose to braves in 10th, fire coach:
  2011-06-09 00:00:00.0
```

Figure 3. User Interfaces of the Proposed Model.

$$\boldsymbol{Precision} = \frac{\boldsymbol{Good\ Articles\ Recommended}}{\boldsymbol{All\ Articles\ Recommended}} \tag{1}$$

Or $$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{Good\ Articles\ Recommended}}{\boldsymbol{All\ Good\ Articles}} \tag{2}$$

Or $$Recall = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$$

$$\boldsymbol{F-Score} = \frac{2(\boldsymbol{Good\ Articles\ Recommended})}{2(\boldsymbol{Good\ Articles\ Recommended}) + \boldsymbol{Good\ Articles} + \boldsymbol{Not\ Good\ Articles}} \tag{3}$$

Or $$F1 - Score = \frac{\mathrm{2TP}}{\mathrm{2TP+FP+FN}}$$

We have done the performance evaluation of our model in two ways. One is the experimentation to compares each user with each other and the other is the experimentation of the system to compare the performance of categories of users which consists of many users.

i. **Experimentation for individual user similarity**

For the individual users we evaluated by taking 16 which means 10% of active user for testing dataset from dataset which contains 160 active users. Then we remove the news rated by these 16 users and we register each of these 16 users as a new user and we run our new model to recommend the news for the users. And then we compared the previous news rated by the user with these new results. Then, we calculated the Precision, Recall and F1-score values as the formula we discussed in the previous section. According to this experimentation, results of the work is explained in the Table 2.

Table 2. Experimentation value based on each user similarity.

User ID	Precision	Recall	F1-Score
587	0.740741	0.40404	0.522876
888	0.787037	0.429293	0.555556
1263	0.777778	0.424242	0.54902
1932	0.787037	0.429293	0.555556
1973	0.787037	0.425	0.551948
2251	0.731481	0.39899	0.51634
2604	0.648148	0.443038	0.526316
2668	0.583333	0.398734	0.473684
2733	0.481481	0.396947	0.435146
3148	0.611111	0.420382	0.498113
3832	0.62963	0.427673	0.509363
4364	0.564815	0.388535	0.460377
4487	0.62037	0.421384	0.501873
4122	0.601852	0.414013	0.490566
4210	0.537037	0.367089	0.43609
2360	1	0.605263	0.754098
Average	**0.680556**	**0.42462**	**0.521058**

From the evaluation results, we have the accuracy of the news to each individual user with the values of 68.05% of average Precision, 42.46% of average recall and 52.1% of average of F1_score values as shown on Figure 4.

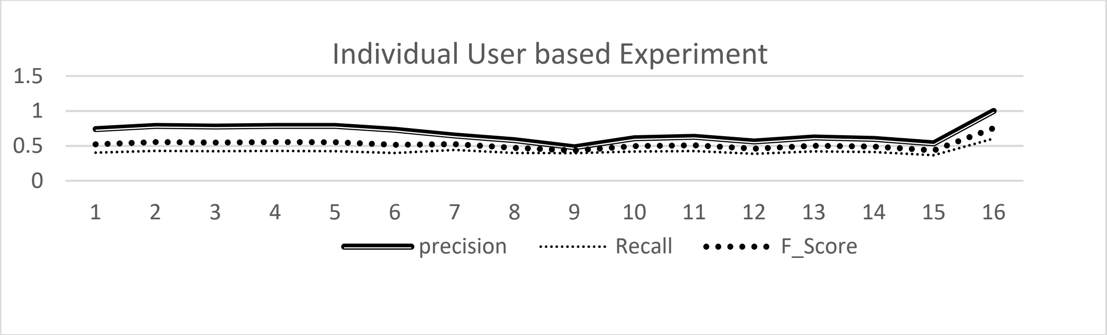


Figure 4. Experimentation value based on individual user similarity.

ii. Experimentation by user category-based similarity

The other way we evaluated our performance is based on the user category. We took 5 user groups which of 10% of user groups out of user categories we have in our dataset. Then, we recommend some of new users similar to the category selected and we compare the accuracy performance by comparing the actual recommendation with the recommended for that category. According to this experimentation, results of the work is explained is shown in Table 3.

Table 3. Experimentation values for category-based user similarity.

category Number	Precision	Recall	F1-Score
2	0.944444	0.408	0.569832
5	0.891304	0.366071	0.518987

16	0.944444	0.409639	0.571429
22	0.944444	0.408	0.569832
31	0.962963	0.421053	0.585915
Average	**0.93752**	**0.402553**	**0.563199**

Based on this experiment we have the accuracy values of that performs the average precision of 93.75%, average recall of 40.25% and average F1-score of 56.31% and the average of the precision on this experiment and it is shown on Figure 5.

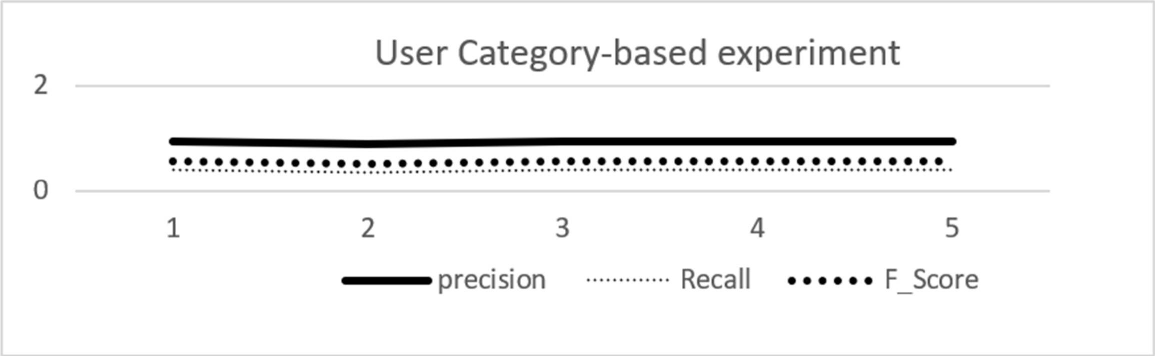


Figure 5. Experimentation value based on user category similarity.

4.2. Discussion

The result of our experiment in this work contains two experimentation way and each has different values as discussed in Section 4.1. Finally, we analysed that the recommendation performance is different as obtained in two ways of our evaluation methods. The recommendation accuracy result is different according to the similarity of the users with category based and individual user similarity. According to the two experiments results, we analyzed that the more accurate recommendation is done for the users in the category which performs the precision of 93.75 %, Recall of 40.25% and F1-score of 56.31% rather than individual user recommendation which performs 68.05% of average Precision, 42.46% of average recall and 52.1% of average of F1_score. This is done because of the category-based recommendation contains more related news regarding to the users in that category more than the news recommended for individual user similarity. According to this value the Recall in the individual performs less performance than in category-based. The reason behind this result is, the finding of the good recommendation result numbers from many users in the category. So, if the good news articles recommended are many then, the Recall value will become less.

5. Conclusions

In this work, User Demographic Data with Hybrid news Recommendation system which considers popular news is proposed. This hybrid recommender scheme combines content-based and collaborative approach with user demographic information. In our case, the content-based uses the similar user in the same category rate for mostly rated categories of news. The collaborative filtering uses the user similarity to check the news rated with the similar groups of users. The similarity of users is done based on the demographic information user's registers in to the system at start time for the system. The user demographic data such as gender, age, and profession and user id in the system helps to identify similar users from existing users by categorizing. In addition to these approaches, the popular news mostly read by many users and rated with high-rate value is filtered. Finally, the hybrid approach we proposed combines these results and provide recommendation by ranking in time recency order and recommended relevant news articles for the new users. In addition to this, the large data processing or the scalability problem is handled by categorizing the news dataset and user demographic data set.

In order to evaluate the performance of the proposed news recommender scheme, an extensive experiment is conducted. Moreover, the performance is evaluated using Precision, Recall and F1-Score of information accuracy metrics. The proposed model achieved an efficient performance result towards Precision, Recall and F1-Score of information accuracy metrics. The experiment results demonstrate that User Demographic Data with Hybrid news Recommendation system achieves a satisfactory performance result. So, it is indeed, there is a satisfactory performance improvement. However, the performance assessment should be conducted using some real dataset to take into account different considered scenarios, to conclude the comprehensive effectiveness of the proposed model.

6. Future Work

The effectiveness of a given news recommender system is determined by the features a given approach utilizes, either features of the news itself or information about users, both implicit and explicit data about the users' online behaviours. In regards to information about users, one aspect which requires further investigation is combining two or more user demographic data, to improve user similarity so that suggestion of relevant news article for new user would be better improved. Moreover, the performance assessment should be conducted using some real dataset taking into account different considered scenarios, due to the heterogeneity of users' online behaviours. In addition, although the availability of large-scale dataset is rare, supervised machining learning approaches can be another line of future work to streamline the design of an effective recommender schemes to enhance the performance towards addressing new user cold-start problems.

References

1. C. Shahabi and Y. Chen, "Web Information Personalization: Challenges and Approaches," i*n: Bianchi-Berthouze N. (eds) Databases in Networked Information Systems*, vol. 95, pp. 1–10, 2003.
2. J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," *Knowledge-Based System*, vol. 46, pp. 109–132, 2013.
3. J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, "Recommender System Application Developments: A Survey," *Decision Support Systems*, Vol. 74, pp. 1–38, 2015.
4. Le Hoang Son, "Dealing with the new user cold-start problem in recommender systems: A comparative review," *in: press. Information Systems*, 2014.
5. Gediminas Adomavicious and Alexander Tuzhilin, "Towards the Next Generations of Recommender Systems: A Survey of the State-of-the-Art and Possible Extension," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no 6, pp: 734 - 749, 2005.
6. https://en.wikipedia.org/wiki/News.
7. https://en.wikipedia.org/wiki/Online_newspaper.
8. Mansi Sood, Harmeet Kaur, "Survey on News Recommendation," *International Journal of Advanced Research in Electrical Electronics and Instrumentation Engineering*, Vol. 3, Issue 6, 2014.
9. F. Garcin, C. Dimitrakakis, and B. Faltings, "Personalized news recommendation with context trees," *in: The Proceedings of the 7th ACM conference on Recommender Systems*, pp. 105-112, 2013.
10. P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "Grouplens: an open architecture for collaborative filtering of netnews", *CSCW, ACM,* pp 175–186, 1994.
11. K. G. Saranya and G. S. Sadhasivam, "A personalized online news recommendation system," *International Journal of Computer Applications*, pp 6–14, 2012.
12. L. Li, D. Wang, T. Li, D. Knox, and B. Padmanabhan, "Scene: A scalable two-stage personalized news recommendation system," *in: Proc. The 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 125-134, 2011.
13. M. Balabanović and Y. Shoham, "Fab: Content-based, collaborative recommendation," *Communications of the ACM*, vol. 40, pp. 66-72, 1997.
14. L. Zheng, L. Li, W. Hong, and T. Li, "Penetrate: Personalized news recommendation using ensemble hierarchical clustering," *Expert Systems with Applications*, vol. 40, Issue. 56, 2012.

15. L. Li, L. Zheng, and T. Li, "Logo: A long-short user interest integration in personalized news recommendation," *in: Proc. the fifth ACM Conference on Recommender Systems*, pp. 317-320, 2011.
16. N. Zheng, L. Qiudan, L. Shengcai, Z. Leiming, "Which photo groups should I choose? A comparative study of recommendation algorithms in Flickr," *Journal of Information Science*, vol. 36 no. 6, pp. 732–750, 2010.
17. E. Brynjolfsson, Y.J. Hu, M.D. Smith, "Consumer surplus in the digital economy: estimating the value of increased product variety at online booksellers," *Forthcoming in Management Science*, vol. 49, no. 11, pp. 1580–1596 ,2003.
18. Hofmann, T., "Latent semantic models for collaborative filtering," *ACM Trans. Info. System*, vol. 22(1):89-115, 2004
19. D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, vol. 35, pp. 61-70, 1992.
20. P. Cotter and B. Smyth, "Ptv: Intelligent personalized tv guides," *in: AAAI/IAAI*, pp. 957-964, 2000.
21. M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes and M. Sartin, "Combining content-based and collaborative filters in an online newspaper," *in: Proc. ACM SIGIR Workshop on Recommender Systems*, 1999.
22. Tranos Zuva, Sunday O. Ojo, Seleman M. Ngwira, and Keneilwe Zuva, "A Survey of Recommender Systems Techniques, Challenges and Evaluation Metrics," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 11, 2012
23. Marko Tkal c., Matev z. Kunaver, Andrej Ko. Sir, and Jurij Tasi c., "Addressing the New User Problem with a Personality Based User Similarity Measure," no. 1, 2010.
24. Park, S.T., Pennock, D. M., Madani, O., Good, N., and DeCoste, D., "Naive filter bots for robust cold-start recommendations," *in: Proceedings of KDD ACM*, vol. 06, 2006.
25. Salter, J. and Antonopoulos, N., "Cinema Screen recommender agent: combining collaborative and content-based filtering," *IEEE Intelligent Systems*, vol. 21, pp. 35-41, 2006.
26. Schein, A. I., Popescul, A., Ungar, L. H., and Pennock, D. M., "Methods and metrics for cold-start recommendations," i*n: Proceedings of SIGIR ACM*, vol. 02, pp. 253-260, 2002.
27. Pazzani, M., "A framework for Collaborative, Content Based and Demographic Filtering," *Artificial Intelligence Review*, pp. 393-408, 1999.
28. Z. Lu, Z. Dou, J. Lian, X. Xie, and Q. Yang, "Content-based Collaborative Filtering for News Topic Recommendation," *in: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15)*, pp. 217-223, 2015.
29. Hee-Geun Yoon, Hyun-Je Song, Seong-Bae Park, and Kweon Yang Kim, "A Personalized News Recommendation using User Location and News Contents," *Applied Mathematics & Information Sciences*, vol. 9, No. 2, 2015.
30. A. Darvishy, H. Ibrahim, A. Mustapha, and F. Sidi, "New Attributes for Neighbourhood-based Collaborative Filtering in News Recommendation," *Journal of Emerging Technologies in Web Intelligence*, vol. 7, no. 1, pp. 13–19, 2015.
31. M. Sunitha and T. Adilakshmi, "Recommender Systems to Address New User Cold-Start Problem with User Side Information," *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 18, no. 2, pp. 17–23, 2016.
32. F. O. Isinkaye, Y. O. Folajimi, B. A. Ojokoh, "Recommendation systems: Principles, Methods and Evaluation," *Egyptian Informatics Journal*, vol. 16, pp. 261-273, 2015.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.