JIMMA UNIVERSITY
INSTITUTE OF TECHNOLOGY
SCHOOL OF COMPUTING

News Recommendation System for New User using Hybrid Approach with Demographic Data

Zerihun Olana Asefa

A THESIS SUBMITTED TO THE SCHOOL OF COMPUTING OF JIMMA UNIVERSITY IN PARTIAL FULFILMENT FOR THE DEGREE OF MASTERS OF SCIENCE IN INFORMATION TEECHNOLOGY

Jimma, Ethiopia

November 15, 2017

JIMMA UNIVERSITY
INSTITUTE OF TECHNOLOGY
SCHOOL OF COMPUTING

Zerihun Olana Asefa

Advisors:

Advisor: *Melita Luke (PhD)*

Co-Advisor: *Behailu Shewandagn (M.Sc.)*

This is to certify that the thesis prepared by *Zerihun Olana*, titled: *News Recommendation System for New User using Hybrid Approach with Demographic Data* and submitted in partial fulfilment of the requirements for the Degree of Master of Science in Information Technology complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Approved by the Examining Committee:

Name	Signature	Date
1. Dr. Melita Luke, Advisor:	_____	_____
2. Mr. Behailu Shewandagn. , Co-Advisor:	_____	_____
3. External Examiner:_____	_____	_____
4. Internal Examiner:_____	_____	_____

Dedication

This thesis work is dedicated to my father Olana Asefa and my mother Takelech Yilma with all my sisters and my brothers.

Acknowledgments

When I express thanks to my advisor Dr. Melita Luke I am with great honour for her very effective inspiration, perceptive comments and strong guidance at every time whenever I got doubt. I would also like to thanks my co-advisor Mr Behailu Shewandagn for providing me good guidance and feedback. In addition to my main and co- advisors I would like to thank the Computing staff members Mr Debela Tesfaye and Mr Getachew Mamo for their generous support from the beginning to the end of my thesis works.

As well, I would like to express gratefulness to the following people who helped me in the making of this work with their important help; especially Mr Andualem Chekol for his great motivation and general support at any time. And also, Fikadu Woyessa, Kelemu Deres, workineh Tessema, Berhanu Anbase and I am appreciative to many of my university colleagues to support me during my this Msc thesis.

I must express my appreciation and thankful to my parents and to all my family members for providing me with consistent encouragement in all of my study and through the process of thesis work. Thank you.

Contents

List of Figures

List of Tables

Acronyms

CBF: Content Based Filtering

CF: Collaborative Filtering

IBCF: Item Based Collaborative Filtering

KBF: Knowledge Based Filtering

NRS: News Recommender System

RS: Recommender System

UCBF: User Based Collaborative Filtering

Abstract

Recommender systems are fundamental solutions to information overload on the web due to the availability of multitude information sources. This system filters and presents relevant information to customer and/ or online users, a small subset of items that she/he is most likely to be interested in. The application of Recommendation schemes ranges from entertainment application (i.e., movies, music) to online newspapers information sources to recommend for the users based on their preferences. News recommendation scheme 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 existing news recommendation scheme is limited during a scenario where information about a user or information about set of users in the system is unavailable. This leads to the occurrence of new user cold start problem. Therefore, the main objective of the study is: designing news recommender system using hybrid approaches to address new user cold start problem to ease and suggest more related news article for new users.

In order to achieve the aforementioned objective, User Demographic Information (Data) with Hybrid Recommendation system is proposed. This hybrid recommender scheme combines content-based and collaborative filtering approach with demographic filtering.

To evaluate the effectiveness of the proposed model, an extensive experiment is conducted using news articles dataset with user rating value and user demographic data. The performance of the proposed model evaluated using precision, Recall and F1-Score metrics which support the effectiveness of the proposed model. The proposed model performance is done by two ways of experiment. So, the performance of proposed model performs around 68.05% of Precision, 42.46% of Recall and 52.1% of average of 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 cluster which is better than the first experiment.

Keywords: ***News Recommendation System, Clustering, Cold start Problem, Hybrid Approach, Demographic information, New Users, Popular News.***

Chapter One

Introduction

1.1 Motivation

Today the information available on the Internet is huge and uncontrolled for users in quantity. The most important information needed for user's day-to-day life is changed into the computerized data available through Internet. The users need only necessary information for their purposes from Internet documents. To do this it needs more specialized systems rather than traditional Information Retrievals systems to provide the only necessary and interesting information to users.

Information Retrieval system is the popular system used for providing information for users from many sources available on WWW or any corpus. The system is used to cover every information source available on Internet and it will provide relevant documents for users. The returned information may not be the necessitated information for users. So, it is good to provide only the appropriate documents or information for each user depending on their interest. The best solution proposed for this problem is Recommendation system.

Recommendation system is the system which depends on the history of users and the information or documents accessed by users to make future recommendation. Recommending or suggesting the users is a good solution to save the users from the overwhelming information which are not part of their interest. The users' interests are different depending on the domains of the information they need to access from the available information on the Internet. For example, users may search for entertainment information, for academic, for news or for commercial products. The users that want news information should get only information available from news source [1], [2], [3].

The recommendation system uses basically the contents of the items and the users profile to provide the information. Thus, the items must be known by the system and also the user's information is quite important to predict the user's interest. Then the system will compute the relation of the items to be provided with the users from the users past history.

And also 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.

Recommendation system uses different techniques to solve the problem with traditional information retrieval system. These techniques use the information of the users who has to be recommended and the products or items that system will recommend. So, the contents of the items and the profiles of users are the important information in the recommendation system [1].

The motivation for this work is to recommend news for any reader even for the users with no history in the Recommendation system. News recommendation is the domain of recommendation system which we are motivated for this work. The work is used to suggest news to online readers according to their interest of news from available online newspapers.

The news released daily in our world through the Internet is huge and it is not convenient for the readers to access the news specifically needed by them because of the availability of several newspapers. So, rather than searching for all newspapers and news it is good to be recommended by the system only the interested and related news for readers depending on user's profile history which is possible with the News Recommender system. This system will provide only the relevant information for the readers of news depending on their interest of news categories. This news recommendation system is used to help the users on their interest of news and to provide the updated news for the users based on their interest.

Many works have been done previously focusing on other similar areas like recommending popular news, fresh news and topic based news. Still there is problem on providing the relevant information for new readers or new users who have no history of data in the recommendation system. Our motivation for this study is to recommend the relevant information for new users which have no history in the system. The recommendation system needs the history data of both the item to be recommended and the user's pattern of interest to make personalized recommendation system [3].

Our proposed system will use the hybrid of Content-based and Collaborative Filtering approaches for news recommender to predict the recommended news for readers.

1.2 Statement of the problem

The availability of news articles from the World Wide Web is not sufficient enough for news reader's to access news based on their preferences, due to the availability of enormous amount of newspapers with different categories of news available on the web in the Internet. Moreover, users use each search query in order to access news every time on different sources of newspapers. However, it is difficult to get more relevant news out of all available newspapers. The retrieved news using the traditional information retrieval mechanism will provide all related news with the reader's query without taking into account user's interest. For example, some reader's need only specific field of news which is more related with their profession. Whereas, others users want to read more general field (i.e., sports or entertainment). Therefore, recommender schemes in general, and news recommendation system in particularly plays a crucial role to fill this gap.

Most news recommendation system used user's data and content of the news itself, to suggest and recommend relevant news articles for a particular readers based on the interests he/she has [3]. This user data is divided in to two, explicit data which is collected by reader's direct activities (i.e. when users give feedbacks directly to the news articles while she/ he is reading) and implicit data which the system collects by following 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 enough in order to suggest and recommend relevant news articles to the readers in the scenario where a new users joins the system [1], [3], [4].

The lack of available user data in the situation where new user/ users joins the system, limits the effectiveness of news recommender system to suggest relevant news to this kind of users. Due to this insufficient user data, there is an occurrence of new user cold-start problem. The new user cold start problem is the common problem in the recommendation system because of lack of information of new users in the system [16]. Furthermore, users gets irrelevant news feeds that are not related to their preferences.

In this thesis, we investigate and examine the following research questions.

1. How to model user demographic with hybrid approach?
2. How to figure out this impact on news recommender scheme towards performance metrics such as, Precision, Recall and F1-Score?

3. How to model a news recommender system for new user?

1.3 Objectives

1.3.1 Main Objectives

The main objective of this thesis is to propose the News Recommender system for online newspapers to solve the problem with new user cold start problem which will provide the more related news article for readers by using hybrid approaches.

1.3.2 Specific Objectives

The following specific objectives are formulated from the general objective.

- To provide interesting news recommendation for new users
- To solve the scalability problem using clustering algorithm
- To recommend popular news
- To cluster user based on their demographic information.
- To develop the hybrid approach for News recommender system with demographic data.
- To evaluate the effectiveness of the recommended news articles.

1.4 Methodology

The methods or techniques to be applied for this study to achieve each specific objective is done by reviewing different related papers to our study to get what has been done before and to identify the research gaps to be performed by this proposed system. The other methodology we will use in our study is the method of modelling and implementing the architecture. The data collection and data processing is another method we consider in our work. And the evaluation methods used for evaluating the performance of the system and comparison of the result with existing system is another main issue that is performed.

1.4.1 Literature Review

There are many works done which are more related to our work. So, different works done before are reviewed with their difference in algorithm design and architecture. We reviewed research articles that used some similar algorithms and their approaches for identifying and designing their model.

1.4.2 System Architecture

The system architecture is the core thing for the work we implemented. The modelling of the system is done by considering the objectives of this study with the logical order of components of this work. The techniques used to implement an appropriate algorithm for this work is designed and discussed for describing our works clearly.

1.4.3 Implementation Tools

The tools used in our work consider all of process we use in the study. So, the tools we used for data preparation, for implementing of the proposed model algorithm and designing the model. To store the collected and processed data from news dataset, we used MySQL. We also used Microsoft Visio for architecture design and Java programming to process and provide the recommended news based on reader's interest.

1.4.4 Data collection

The dataset used for evaluation of the proposed system is collected from the news dataset which is available on the Internet and published is English language from popular GitHub websites which collects data from different websites and search engines. The collected data from the website consists of information of news articles features. Another dataset we used is the user demographic dataset prepared by ourselves and the rating dataset collected by interacting with the users and the news articles.

1.4.5 Evaluation methods

The collection of result from the system, observing the result, statistically analysing them to evaluate the system performance is necessary. Analysing the proposed work with performance metrics is used to identify the contribution achieved in the study. It is used to get the performance of the proposed system and the performance of the approaches used. The system performance is evaluated by using the accuracy metrics (Precision, Recall and F1-Score) on the items retrieved against user's interest.

The collection of results from the system, statistically analysing the results obtained is done by using tables and charts.

1.5 Scope

The new user's data used for the proposed system is the only demographic information of users when the registration is done for the website. And the system is not functional for the users not registered to the system.

1.6 Significance of the study

The study will develop the system that will return the most relevant news articles for the new readers from many available online newspapers. The returned news will be the interesting news articles for the users. The user profile of users for that particular news websites will be provided to the recommender system.

As our proposed system is to fix the problem of new user interest of news, the system will fix it and recommends news for the user according to user groups and their patterns of interest. In addition it recommends the popular news by filtering the most read articles by many users and get the highest rate value. And the users should recommend with only the categories he/she is more interested with. For example, if he/she is interested by entertainment news, he/she should have to get from this category and the same for others news categories. The system returns the news to all new users and existing users which is the category-based (i.e. the news to be recommended/predicted will be the category of news that users want to read).

1.7 Organization of the Thesis

The rest of the document is organized as follows Chapter two presents Literature Review, Chapter three presents Related work, Chapter four presents proposed system, Chapter five presents Experimentation Result, Discussion and Evaluation work, Chapter six presents Conclusion and Future work suggestions.

Chapter Two

Literature Review

2.1 Introduction

This chapter presents the state of the art in the Recommendation system and News Recommendation System with overview of their components and the techniques developed. This part of study enlightens briefly on some of the basic concepts of the Recommendation system and the techniques applicable with the system.

2.2 Overview of Recommendation System (RS)

The ubiquity of the web brings an explosive increase of accessible information. Recommendation Systems have emerged as an intelligent information filtering tool to help users effectively identify information of items of interest by users from such an overwhelming set of choices and provide personalized services [7]. Therefore, recommendation systems are considered to be a critical tool for enhancing sales in e-commerce websites [8], [9].

Due to incredible and increasing growth of information sources available online, the World Wide Web is witnessing a rising demand of intelligent systems that can guide users to find relevant information for them. Search engines can help to resolve this problem to an extent, particularly if users are looking for some specific data that can be formulated as a formal query. However, in many cases, users may not even know what to look for. Often this is the case with items like news, movies etc., where users are browsing for things that might match up their interest areas. In such cases, it is better to present open recommendations to users based on their interests [4].

The Internet has no shortage of content for user's interest and contents to be provided for users. However, the challenge is how to find the more interesting content for users,

something that will answer user's current information needs or something that users would like to read, listen or watch. In such cases, it is important to present recommendations to a user based on his/her interests as demonstrated by his/her past activity. A recommendation system can be used to suggest customized information according to user preferences.

With the explosion of service based web application like online news, shopping, bidding, libraries great amount of information is available. Due to this information overload problem, to find right thing is a tedious task for the user. With the dramatically fast and explosive growth of knowledge on the market over the Internet, World Wide Web has become a robust platform to store, spread and retrieve data likewise to mine helpful data. Due to this large amount of information, finding interesting information is a tedious and time consuming task for the user. This reason brings recommendation system into light to assist the users to satisfy their interest. The recommendation Systems is a software tools and techniques that deal with information overload by providing interesting suggestions and recommendations to users [10].

2.2.1 Common Techniques of Recommendation System

The recommendation system uses several approaches for providing the best results in the form of recommendations for an individual who is looking for favourites from an online recommendation system. In the present time the following approaches are being used, namely, Collaborative filtering (CF), Content-Based filtering (CBF), Knowledge-based filtering (KBF) and Hybrid Filtering (HF) [11]. Figure 2.1 [60] shows the different classifications of techniques in RS.

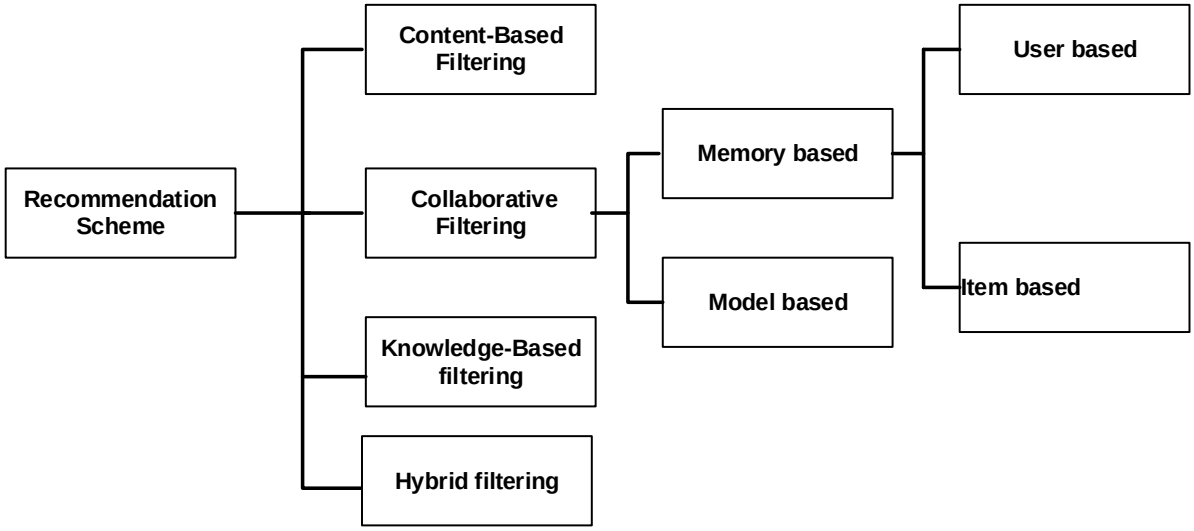

Figure 2.1 Recommendation System Taxonomies

A. Content-based Filtering: Content-based recommender offers recommendations based on target user ratings and items associated features. It assumes that user will rate items having similarly to their interest. This approach recommends new items having similar features to the items which have been rated by the user.
B. Collaborative Filtering: Collaborative recommender provides recommendations based on users' similarity, it assumes that users with similar tastes will rate items similarly. It attempts to find users having similar rating history to the target user (user who requires recommendations), building a neighbourhood from which the recommended items are generated.
C. Knowledge-Based Filtering: Knowledge based recommendation systems makes use of knowledge structure to make inference about the user needs and preferences. Such kind of recommender systems have knowledge about what kind of items are liked by a user based upon user profile information and context-related information. So, a relationship can be established between user needs and relevant recommendation out of available collection of items for that user.
D. Hybrid approaches: In this approach the recommendation schemes use the combination of two or more techniques in order to take advantages of each approaches so that the hybrid approaches overcome the problem with individual approaches.

In our work the hybrid approaches is used having both collaborative filtering and content-based filtering techniques. The collaborative Filtering is used to filter the articles read and rated by similar users to estimate the future of the user. The second one, the content-based was used to filter the news contents based on news attributes like time and category. The hybrid method of both content-based and collaborative filtering is the most widely used techniques in News Recommendation System.

2.2.2 Challenges of Recommendation System

The recommendation system by itself has many challenges to provide better performance. Those challenges are under research to be solved by many researchers using different algorithms. The most common challenges of Recommendation system are sparsity problem, cold start problem, scalability problem and over-specialization problem as studied in survey of [2], [5], [6].

Sparsity: Sparsity is the problem that have influence on the accuracy of the system. It is caused by lack of users rating data. Since many users are not willing to rate the items they accessed this problem will occur. If no rating data, it is difficult to know the user's history for recommending in the future.

Cold start problem: It is the common problem in recommendation system which happen because of lack of information of users need and the items needed to be recommended by the system.

Scalability: The problem of many data to be processed while the recommendation is done by the system. If the data used to provide recommendation is huge like in the commerce system which deals with many users and many items with their rating history.

Over-specialization: Over-specialization is the problem in the items based recommendation for recommending the item based on their similarity. If more items are similar, then the items to be recommended from this group of items will become more and more which follows over-specialization.

2.3 News Recommender System (NRS)

News is current information which is provided by journalists via 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 the information from around the world. News are the information most accessed by many readers [12].

News articles are published and reach readers on news website for the online users and on print paper for printed paper readers. The online newspaper is the online standard of news published on Internet in series, which contains information about current events and other informative articles on different categories, such as politics, sports, arts and advertising [13]. Today there are many and different readers willing to read different newspapers available on the Internet by using different sources of online newspapers. And also, there are many sources of newspapers to provide the news for these users. Many newspapers are specified for single categories (i.e. sports, Arts, Vacancy Announcement, Business, Politics, etc.). Even though it is categorized, users want to read only some news provided by different news provider for their interest.

The online newspapers provide the important information for the readers on the since news is the updated and contains the current actions in the world. Because in today's life most users are connected to the Internet, they can access and get any new activities from around world. In addition, newspapers provide news for readers should have to be substance information. So, the readers should have to identify the news functionalities in many aspects. The correct and the useful news should have the following elements according to [13].

1. Timeliness: News is the information which transfers new things for the readers. So, the information should be new at the time published and its timeliness could be different from one publishers to others.
2. Impact: The impact of one news or events happened may have different impact to other side of people. So, if the events happened in one area of the world have an impact on other rest of world it is good to take care for them.

3. Proximity: If some accident happened is more nearest to the other it is proximity for that areas people for the consequence follows.
4. Controversy: Anything that is connected with conflicts, arguments, charges and counter-charges, fights and tension becomes news for the readers.
5. Prominence: This is about the more popular people events to become more news than other ordinary people.
6. Currency: The current events is interesting since most peoples are attracted by the current activity.
7. Oddity: Unusual thing, extraordinary and unexpected events will create the more news with the public.
8. Emotion: The news with story of people which can be bad or good history will become more interesting to the readers.
9. Usefulness: The news will help the readers for what will happen and happening. Not only this but also what happened somewhere else can be used to protect themselves or to be ready for similar events.
10. Educational value: In most newspapers there are many news about educational course and job opportunities. So, readers will be helped by this useful information.

Online users access many news to achieve their interest and to satisfy their need. They have to try in different ways based on the news they are interest with it. The news of articles information about various issues is available for the users. Therefore, news recommendation system helps in narrowing down this information and managing this information by estimating the choice of a user. This is done by analysing online news which is a key enabling technology and assists users for various items.

News Recommendation systems have evolved as an answer to information overload problem prevalent with online news readers, while users looking for relevant news information out of a huge premise of news contents 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 recommendation systems where these systems are used to suggest news to the users that match their interests and personal preferences.

Since many information at different time is published on the news website there will be the big of information on these websites. But the user needs the recent news and the system.

2.3.1 Open Issues of 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's that news should have to include in its system there are some challenges when recommending relevant news articles for readers. So, to consider such challenges and making recommendation 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 challenges studied in the survey work of [14], [15], [17], [18], [32] are discussed as follows. Some of them are common for all kinds of recommendation systems are explained earlier.

1. Scalability: The volume of news is large since it is collected from different source of online newspapers with different categories of news within a short period of time.
2. Popularity of news articles: Since the news are going change from time to time dynamically within a period of time according to the readers and frequency of the articles it is difficult to recommend effective information.
3. Occurrence of specific events; many readers want the news articles that consist of updates about specific person, place or objects due to some special liking about the special topic. But the recommender systems will face to easily predict such specialty.
4. Cold start problem: It is the common problem of Recommendation system in many domain of Recommendation system which occurs for the lack of information of new users and new news articles for the news recommendation engine.
5. User profile data sparsity: Most of 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 the series information they read like politics or other serious news.

6. Freshness of news articles: The special characters of news is the information readers feed from it is the fresh and the current activities done. But if the fresh news is not relevant to the users the system will not recommend the breaking news.
7. The structure of news text: When news is published they are not structured format because of the content that it should consider to be published and it is difficult to analyse like other information on the websites.
8. Life time: The news life time is short and it is not convenient to manage it in the recommendation system since the recommendation system needs the past information to recommend for the future.
9. Similarity between news articles does not necessarily reflect their relatedness. This is another challenges in news recommendation system which is not easy to manage because of the news articles might be similar in many words but completely different in their message.

The challenges listed out above are the main ones in news recommendation system and many studies have been done to fix them even though still they are not complete. In our study we propose 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 the related news for the users. So, the readers are going to read the news they will not like to read.

2.3.2 Popular Approaches of News Recommendation System

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

A. 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 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 rating history of current user related to previously used items.

Similarity of items is determined by measuring the similarity in their properties. So, in this type of filtering method there is no dependency on rating records of other users in order to generate preferences for current user. Content-Based systems focus on properties of items. For example, if the user has purchased a book on amazon.com which uses recommendation system then the user starts getting additional preferences for buying books from online book store which includes 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 interest to the user. In this method, news recommendation is done based on 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 a matter of information retrieval (IR), where the content related to the user's preferred choices is regarded as an enquiry, and unrated documents are evaluated on the basis of relevance/similarity to this enquiry [19], [20], [21].

In order to recommend one particular news to a user, the content-based approaches will get the previously rated news based on their category as an example, like (sports, science & technology, health, business) and then the news with highest similarity to user preferences are recommended.

B. Collaborative Filtering (CF)

The other technique which is used in our study is Collaborative filtering. In Collaborative filtering based RS, the user will be suggested items that people with similar interested and preferences liked in the past. In a CF recommendation system, in order to suggest items to user, the collaborative filtering recommendation system looks for the peers of user, i.e., set of users that have similar interest in item. Then, only the items that are most liked by the peers of 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 research community.

In addition, Collaborative filtering (CF), as a kind of personalized recommendation technique, has been widely used in many domains [22, 23]. However, collaborative filtering also suffers from a few issues, for instance, cold start problem, 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 users' history must be available. However, the database is always very sparse, that is, user only rates a small number of items. Up to now, there are many researchers who have focused on the prediction accuracy and proposed some solutions.

The term collaborative filtering (CF) was coined in [9], [24] to refer to personalized recommendations that are generated based on user preferences and item rating matrix. CF-based recommendation systems retain user preferences data and utilize them to identify groups of similar users. Then, user-liked items can be recommended to similar users in the group. These methods recommend news based on previous readings of other users by identifying similar users to the active user [25]. For this reason, CF-based systems are also referred to as social-filtering recommendation systems.

Collaborative filtering (CF) is one of the most successful recommendation technologies. The basic idea behind this method is that it gathers the opinions of other users who share similar interests with a target user (referred to as the "active user") and assists this active user to identify items of interest based on these neighbours' opinions. These social information filtering approaches automate a process of "word-of-mouth" recommendations [26]. Compared to other recommendation technologies (e.g. content-based filtering), CF provides some prominent advantages to information filtering:

(i) the capability to filter items whose content is not easily analysed by automated processes;

(ii) the ability to provide serendipitous recommendations; and

(iii) Support for social factors by taking into account the interests of like-minded users [27, 28].

Consequently, CF has been becoming popular in both academy and industry fields with great speed. Despite the overall success of CF systems, they suffer one serious limitation, namely the cold-start problem [4]. 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 [29].

Collaborative Filtering uses two models to identify the similarity of users and news articles in between them. One is the Neighbourhood model or memory based Collaborative Filtering which uses the data stored on memory to calculate the similarity. The other and the second one is the model based which uses the online method to calculate the similarity between users or items to be recommended. The Memory based or Neighbourhood model is the easy and popular. And it is discussed in detail in the next section of this chapter.

Since collaborative Filtering uses the similarity of user with respective to their items they rated it is based on both user and item. The collaborative Filtering is the techniques use user-based similarity or item-based similarity.

User-based Collaborative Filtering (UBCF)

The assumption of this UBCF is that the People who settled in the past are likely to agree again. And to predict a user's opinion for an item, it uses the opinion of similar users and the similarity between users is decided by looking at their overlap in opinions for other items. User-based Collaborative Filtering has been investigated in several works. And it is one of the widely used in recommendation technologies, remaining 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 on other items in the future [24]. 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)

The similarity between items is 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 calculate difference in ratings for the two items and finally, it takes the average of this difference over the users. For example, if the news in similar category rated by different readers should be considered as similar items.

C. Hybrid Approach

In this study, 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 content-based approach or collaborative filtering with demographic characteristics based recommendation approach [11].

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 [30], [31].

There are many hybrid methods developed on the basis of traditional collaborative filtering, but which also retain the content-based profile for individual users. Such profiles, instead of co-rated items, are utilized to identify similar users. In [32], [33] study, each user profile is characterized by a vector of weighted words obtained from positive training examples employing the Winnow algorithm. Predictions are made through the application of CF directly to the matrix of user-profiles (in contrast to the user-ratings matrix).

Alternatively, an approach known as FAB [34], [35] makes use of relevance feedback to simultaneously mould a personal filter together with a communal topic filter. This method initially ranks a document according to the topic filter and then transmits it to a user's personal filter. The relevant feedback of the user is then used to alter both the personal filter and the originating topic filter.

2.4 Clustering Algorithm

In this study clustering algorithm is used which is a domain of data mining that had been useful in a different problem, such as pattern recognition, image processing, statistical data analysis and knowledge discovery [36]. It is the technique which was used to group the data based on their closeness to each other for solving the problem of time complexity and memory space while the searching process is ongoing. Clustering is the process or method used to group data according to their similarity to reduce the search complexity of query and the memory used to store the data required. The aim of the clustering is to reduce the searching time of large data and also to reduce the memory space used when request is processed.

Clustering methods can be divided into two main groups: hierarchical and partitioning methods [38], [39].

2.4.1 Partitioning clustering

Partitioning clustering is the method require the number of clusters which will be determined by the user to rearranging each objects place starting from the initial partitioning. In this method of clustering the process needs many times to achieve the better result by using all possible partitions. Because of this it is not easy to clusters the

given large data sets. The most common type is the K-Means and it is discussed as the following [38].

K-means algorithm

This algorithm has the objective of classifying a set of *n* contexts into *k* clusters, based on the closeness to the cluster centres. The closeness to cluster centres is measured by the use of a Euclidean distance algorithm. K-means is an iterative clustering algorithm in which items are moved among sets of clusters until the desired set is reached. A high degree of similarity among senses in clusters is obtained, while a high degree of dissimilarity among senses in different clusters achieved simultaneously [40].

2.4.2 Hierarchical clustering

The other common and popular clustering algorithm used to cluster the big data for the effectiveness of searching is Hierarchical clustering. It is the methods create the clustering by segmenting the given data sets in a top-down or bottom-up way [38].

A. Agglomerative hierarchical clustering

The assumption of this method is initially started by representing each objects in its own cluster and then consecutively merging until the preferred cluster result is achieved. This is called the bottom-up way of clustering.

B. Divisive hierarchical clustering

In this case, initially all the objects considered as one cluster and then consecutively each cluster is divided into sub clusters until the wanted result is gained. The top-down way of clustering which starts from the assumption of one cluster into many clustering.

The merging or division of hierarchical clustering is done based on the chosen similarity measures. So, in our work we applied the new user demographic information which is used as an input data was clustered using hierarchical agglomerative approach, then new users were joined to the most similar cluster in the dendrogram as it is stated in [36].

However, the method of clustering used by this study differs from them. Usually, clustering is performed while mining the data. But the method proposed in our work, is to

cluster data while structuring the data model itself. It is used as a data structure for collecting data from the time of implementation and not after collecting user data.

So the hierarchical clustering method proposed here is for data modelling and data management during storage. This is the upcoming research trend which is otherwise termed as attribute clustering to cluster the similar attributes in one cluster of the given data [59].

2.5 New User Cold Start Problem

There are many challenges of Recommendation system as discussed in section 2.2.3, out of them the cold start problem is the common one which happens when user data or item data is scarce. Since a new user, have not data in a system, normally the system cannot get satisfy recommendations. Similar to the new user problem, new items which have not been rated could not be recommended, which is referred to as new item problem. So, the main purpose of our study was focused on the new user problem which is a key issue that determinants the initial success of users.

More accurate recommendations for new users could make these new users stay rather than pushing them to the competitors' sites. 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 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 lack of new item or new user for the system is announced. In recommendation system the cold start problem is common problem.

The new user problem in 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) [41].

Various researchers on hybrid recommendation systems combine both content information and ratings data [42], [43], [44] 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 [45] 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.

2.6 Summary

This chapter contains the discussions of the literature review or the basic concepts of the study area. We have presented and discussed the basic concepts of our study area and some related approaches to our work. The basic concepts of the recommendation system in general and the news recommendation at specific level with their techniques and each component of our works overviews are explained in detail. Some of the discussed concepts are summarized as follows:

Recommendation system is recommending or suggesting the users to save users from providing the overwhelmed information depending on the history of users and the information or documents accessed by users. Thus, the items must be known by the system and the user's information is also the necessary to predict the user's interest. Then the system will compute the relation of the items to be provided with the users from the users past history.

The recommendation system uses several approaches for providing the best results in the form of recommendations for an individual who is looking for resources from an online system. In the present time, Collaborative filtering (CF), Content-Based filtering (CBF), Knowledge-based filtering (KBF) and Hybrid Filtering (HF) are the popular techniques.

News Recommendation system is one domain of recommendation system used to suggest the news readers in online according to the interest of news readers from available online newspapers which are released daily in the world through the Internet. The News Recommendation used to provide the only interested news with users from huge source of news to the readers which are specifically needed by them by searching the related news for readers based on user's profile history.

Clustering algorithm is the one of the methods we apply in our study specifically for clustering big data we used for our works. It is the methods used to collect the same data into the same group based on the specified features of data to be clustered.

New user cold start is the common problems in Recommendation system. This problem occurred usually when the user has no history in the recommendation system. Having the concepts and science discussed in our research area, we will try to dig out better approaches to model the new proposed system.

Chapter Three

Related Work

3.1 Introduction

In this Chapter, we discuss about various approaches towards News Recommendation Scheme. Researchers focus on the new user cold start problem, as one primary issue in recommendation systems, and many studies have been carried out to address new user cold start problem in order to achieve an effective performance [51], [55], [57].

3.2 Hybrid Recommendation System

The performance and effectiveness of hybrid filtering approaches would be improved by combining two or more features. A comprehensive metrics helps the scheme to suggest news articles to the readers based on their preference. The overall News Recommendation Scheme taxonomies are depicted in figure 3.1.

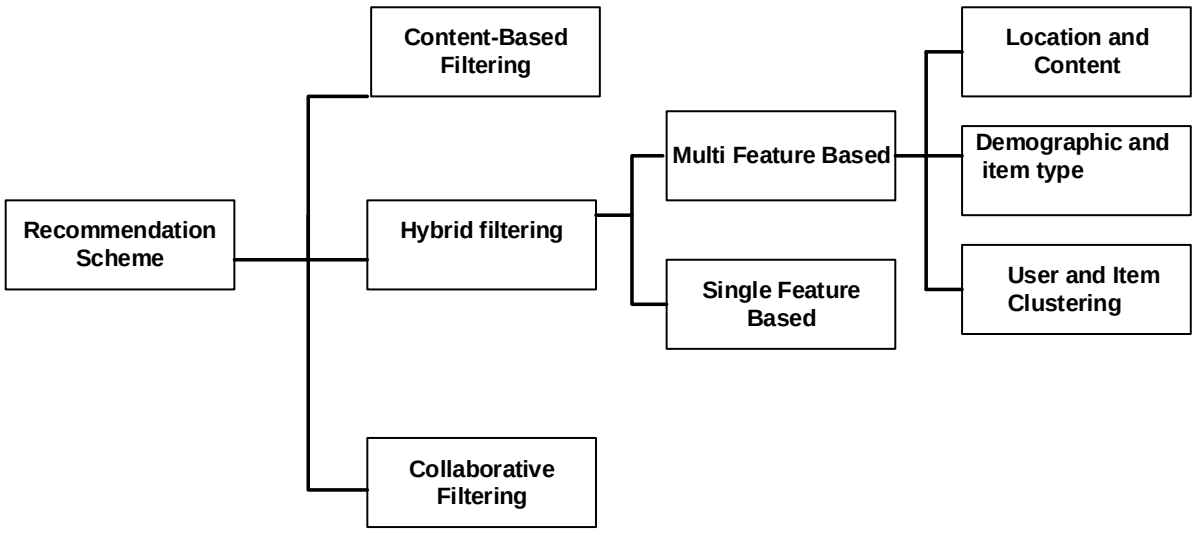

Figure 3.1 News Recommendation System Taxonomies

The study conducted by Zhongqi Lu *et al*. [46], 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 interests of related users. In this experiment and analyses, the performance gains and insights in news topic recommendation in Bing were discussed.

This work brings both Content-based Filtering approach and 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 Collaborative Filtering approach finds similarity in user-item links and clusters the links.

The recent study by Hee-Geun Yoon [47], entitled as a Personalized News Recommendation using User Location and News Contents focus on a specific user based on the preference of the user. With increasing use of hand-held devices, the interests of users are not influenced only by news contents, but also by their location. Therefore, in this study, they propose a novel model to incorporate user location into a user preference for the location-based personalized news recommendation. The proposed model is Spatial Topical Preference Model (STPM). By representing the preference of a user differently according to the location of the user, the model recommends the user appropriate news articles to the user location. For this purpose, they represent geographical topic patterns with a Gaussian distribution. STPM is trained only with the news articles that the user actually reads.

As a result, it shows poor performance, when the user reads just a few news articles. This problem of STPM is compensated for by LDA-based user profile that is not affected by user location. 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 recommendation well, and the combined model outperforms both STPM and LDA. These experimental results prove that the location-based user preference improves the performance of news article recommendation, and the proposed model incorporates the locational information of users into news recommendation effectively.

3.3 Collaborative Filtering

Many researchers have tried to solve the cold start problem in recommendation system in case of many domains like e-commerce, movie and music recommendation by using different approaches. But still there are many problems which are unsolved especially for news readers. We have reviewed many works which are done to solve this problem. Online recommender systems help users more easily and quickly find products that they truly prefer amongst the enormous volume of information available to them. Collaborative filtering (CF) methods, making recommendations based on opinions from "most similar" users, have been widely adopted in various applications.

In the study of R. Hu and P. Pu [48], they encounter one crucial issue remaining to be solved in spite of the overall success of CF systems, namely the cold-start problem. In this study, they propose a method that combines human personality characteristics into the traditional rating-based similarity computation in the framework of user-based collaborative filtering systems with the motivation to make good recommendations for new users who have rated few items.

This technique can be especially useful for recommenders that are embedded in social networks where personality data can be more easily obtained. They first analyse their method in terms of the influence of the parameters such as the number of neighbours and the weight of rating-based similarity. They further compare their method with pure traditional ratings-based similarity in several experimental conditions. The result shows that applying personality information into traditional user-based collaborative filtering systems can efficiently address the new user problem.

S. K. Tiwari and S. K. Shrivastava [49], are studied an approach for Recommender System. In this work the recommender system studied by using Collaborative filtering techniques which is combined with user demographics and items genres. The objective of this study was develop a web based recommender system can be used to suggest customized information according to user preferences. The finding of their study was shown in this paper recommender system generates suggestions for user by combining collaborating filtering on transaction data with rating predicted with user demographics and item similarity by weighted sum of ratings prediction computed from the transaction rating

value, user data and item data. The advantage of this proposed system was that recommender system can deal with cold start in case of new user or new item.

In this study the MovieLens dataset which contains three sets those are rating data, user demographic and item data was used for their evaluation. The three dataset was clustered firstly and the collaborative filtering was applied to combine their results. The system is evaluated using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

L. Safoury and A. Salah [50], have suggested several approaches for building recommender systems which offer items differently to users based on a specific assumption in order to match their interests. Several approaches have been suggested for providing users with recommendations using their rating history most of these approaches suffer from new user problem (cold-start) which is the initial lack of items ratings. This work suggests utilizing new user demographic data to provide recommendations instead of using rating history to avoid cold-start problem. This study presents a framework for evaluating the usage of different demographic attributes, such as age, gender, and occupation, for recommendation generation. The experiment executed using MovieLens dataset to evaluate the performance of the proposed framework.

One research conducted by authors S. Solanki and S. Batra [51], on Recommender System using Collaborative Filtering and Demographic Characteristics of Users which is hybrid recommender system. The study proposes a new hybrid recommender system framework for solving new user cold-start problem by exploiting user demographic characteristics for finding similarity between new user and already existing users in the system. The efficiency of recommender systems can be improved by proposed approach which calculates recommendations for new user by predicting preferences within much smaller cluster rather than from the entire customer base.

The analysis for this work has been done using MovieLens dataset for enhancing the performance of online movie recommendation system. In their work, the recommender systems use variety of data mining techniques and algorithms to identify relevant preferences of items for users in a system out of available millions of choices.

A. Darvishy *et al.* [52], studied on New Attributes for Neighbourhood-based Collaborative Filtering in News Recommendation. 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 recommendation. 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 had 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 in news recommendation with higher precision and F1-score. They conclude that Reading Rate and Hotness in news have a significant impact on personalized news recommendation systems.

Another study conducted by S. J. Gong [53], was a Collaborative Filtering Recommendation Algorithm Based on User Clustering and Item Clustering which is personalized recommendation systems can help people to find interesting things and they are widely used with the development of electronic commerce. Many recommendation systems employ the collaborative filtering technology, which has been proved to be one of the most successful techniques in recommender systems in recent years. With the gradual increase of customers and products in electronic commerce systems, the time consuming nearest neighbour collaborative filtering search of the target customer in the total customer space resulted in the failure of ensuring the real time requirement of recommender system.

In this study, they proposed a personalized recommendation approach joins the user clustering technology and item clustering technology. Users are clustered based on users' ratings on items, and each users cluster has a cluster centre. Based on the similarity between target user and cluster centres, the nearest neighbours of target user can be found and smooth the prediction where necessary. Then, the proposed approach utilizes the item clustering collaborative filtering to produce the recommendations. The recommendation joining user clustering and item clustering collaborative filtering is more scalable and more accurate than the traditional one.

Maddali and Boddu [54], study which entitled as an Implementation of the User-based Collaborative Filtering Algorithm for news recommender system. Collaborative filtering algorithms (CFAs) are most popular recommender systems for collaborating one another to filter the documents they read from the last decade. CFAs have several features that

make them different from other algorithms. The classification accuracy is one among them. A user-based collaborative filtering algorithm is one of the filtering algorithms, known for their simplicity and efficiency. In this work a steady is conducted for its implementation and its efficiency in terms of prediction complexity.

Another works by J. Bobadilla *et al.* [55], on a collaborative filtering approach to mitigate the new user cold start problem which by Knowledge-Based Systems. The new user cold start issue represents a serious problem in recommender systems as it can lead to the loss of new users who decide to stop using the system due to the lack of accuracy in the recommendations received in that first stage in which they have not yet cast a significant number of votes with which to feed the recommender system's collaborative filtering core. For this reason, it is particularly important to design new similarity metrics which provide greater precision in the results offered to users who have cast few votes.

This work presents a new similarity measure perfected using optimization based on neural learning, which exceeds the best results obtained with current metrics. The metric has been tested on the Netflix and Movielens databases, obtaining important improvements in the measures of accuracy, precision and recall when applied to new user cold start situations. The work includes the mathematical formalization describing how to obtain the main quality measures of a recommender system using leave one-out cross validation.

H. Liu *at el.* [56], done on Collaborative filtering which is a new user similarity model to improve the accuracy of collaborative filtering. Collaborative filtering has become one of the most used approaches to provide personalized services for users. The key of this approach is to find similar users or items using user-item rating matrix so that the system can show recommendations for users. However, most approaches related to this approach are based on similarity algorithms, such as cosine, Pearson correlation coefficient, and mean squared difference.

These methods are not much effective, especially in the cold user conditions. This paper presents a new user similarity model to improve the recommendation performance when only few ratings are available to calculate the similarities for each user. The model not only considers the local context information of user ratings, but also the global preference of user behaviour. Experiments on three real data sets are implemented and compared with

many state-of-the-art similarity measures. The results show the superiority of the new similarity model in recommended performance.

In the work Sunitha and Adilakshmi [57], they proposed a new approach to use user's side information in addition to user-item rating matrix to address 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 social matrix to provide recommendations to new users. The experimental results proved that the performance of RS is improved over traditional CF systems.

Generally, 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 exponential growth of Internet, users are facing the problem of Information overloading. Recommender Systems (RS) serve as an indispensable tool to solve information overloading problem.

The Hybrid approach with demographic data is proposed for news recommendation of new user to recommend the new user whom have no history data in the system. This hybrid approach with demographic data uses the combination of both the 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 user with no history since the new user has no history in the system. This proposed model attempts to solve the problem with new user by clustering the new user with the existing user based on their demographic data they register to the system.

3.4 Summary

In summary, the main objective of a news recommender scheme is to suggest a relevant and a timely news articles to the users. However, the performance of recommender system would be affected by different problems. For instance, a new user cold start problem is one of the problem domains where various recommendations approaches are used to address.

News Recommendation Scheme grouped into four major taxonomies. Those are Content based, collaborative, knowledge based and hybrid filtering. However, due to the objective a recommender scheme is proposed, for example, enhancement of its performance and addressing new user cold start problems, hybrid filtering approaches can be further subdivided in to single feature based and multi feature, based on the characteristics of each user and the items, it has in the data set.

The performance and effectiveness of hybrid filtering approaches would be improved by combining two or more features. A comprehensive metrics helps the scheme to a more relevant and accurate news articles suggestion to the readers based on their preference.

The notion that the usage of comprehensive metrics helps the scheme to suggest a relevant and accurate news articles to the users is supported by combining two or more features in the recommendations system. For instance, taking into account users features and the news content itself (i.e., news topic), recommender can suggest given news to the user if there is a similarity towards the topics. Moreover, a user co-location also used to suggest a news articles to the users where by assumption that users in the same location would have a similar interest. In addition, the recommender can include the time stamp of given news when it read by the users in order to recommend recent news to be read by the user.

Although hybrid based recommendation system achieves significant performance improvement, its performance is highly depending on the availability of user's historical data. There are still rooms for further performance enhancement towards the situations where the occurrence new user cold start problem. The Hybrid approach with demographic data news recommendation scheme is proposed. It belongs to hybrid filtering approach.

Our proposed recommender scheme uses both techniques of content- based filtering and collaborative filtering with the demographic data of user. Using of both content- based and collaborative filtering approach is to improve the effectiveness of the system and additional using of user demographic data is to provide the data for new user with no history data to make recommendation more effective. Related work summary of schemes is summarized in Table 3.1.

Table 3.1 Related Work Summary

News recommender	Type of approach	Demographic and News Categories	Location and content	User and News clustering
Zhongqi Lu *et al*., (2015)	Hybrid	X	X	✓
Hee-Geun Yoon (2015)	Hybrid	X	✓	X
Asghar Darvishy *et al.,* (2015)	Collaborative	X	X	✓
Sunitha and Adilakshmi (2015)	Collaborative	X	X	✓
Proposed Hybrid approach with User Demographic Data	Hybrid	✓	X	✓

✓ satisfies the condition

× not satisfies the conditions

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 recommendation since the users in similar location should have similar interest of news. And also, the news time is used to provide recommendation by considering the hot news should be read by any readers. But all this work didn't consider the problem of new readers whose have no history data in the recommendation system. So, we consider the new reader problem as our main problem and we will fix it in this study.

There are many works done to solve the new user cold start problem for another domain of information such as e-commerce, movies information and music etc. The data and the evaluation method that they have used for their study is different from our study as our goal is different. Even though it is similar on main problem, it is different on the objective of the study as our objective is for the more relatedness of information and the study done is to evaluate the rate value prediction. Therefore, the intention of this study is to develop a prototype which recommend online news for the new readers and evaluate the works based on the usage information.

Chapter Four

Proposed Model

4.1 Overview

In this chapter, we outline the problem that we're trying to solve and also design our approaches to hold the problem and how we are going to accomplish each specific objective to achieve the main objective. The data included in the proposed system show how system performs and the approaches we used with their correct flow were outlined and discussed as follows.

4.2 Approaches of the proposed system

Our work proposes the hybrid approach including the features of both Content-Based Filtering and Collaborative Filtering approaches with user demographic information. The hybrid approach in our work and tries to improve the shortcomings of both approaches to provide active recommendations to news readers by solving new user cold start problem.

Content-based Filtering approach relies on the similarity of news content and clusters the news based on the news categories, while Collaborative Filtering approach finds similarity in user and news and clusters the link between users and news.

4.2.1 Content-based Filtering (CBF)

CBF is one of the techniques commonly used in NRS and other RS to make recommendation based on the contents of the items or the information needed to be recommended by using similarity of the items content. When we say similar content, it means that the items which have similar types of contents will be assumed as similar items. So, in the news case of our work, the news with similar category is considered as similar news to be recommended for the readers of the same news categories. For example, the users read from the sport categories is suggested other news from sport categories. In our works, we consider the news articles rated by users to recommend and so the users who rated news from similar category is considered to be recommended the news from that category with highest rating values.

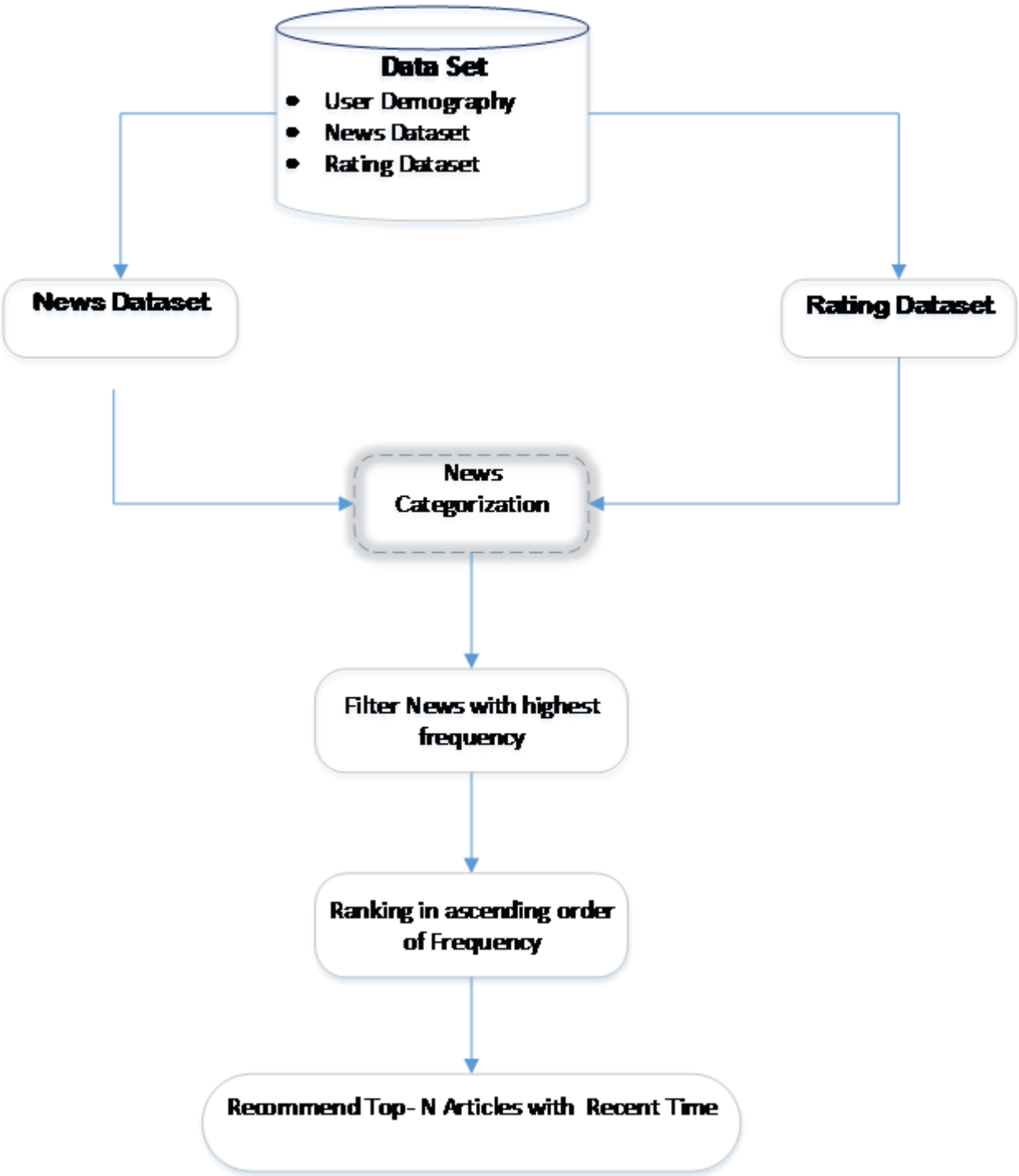

Figure 4.1 content-based Filtering

From the cluster of users done based on user demographic information filtering, the news rated by the users under the cluster should be filtered. Then, from the news filtered the categories of the filtered news is identified and the category which has a highest number of news articles rated are considered as the category that this cluster of users liked. Then, our system will generate the top news from this category by considering the recency of the news and the rating values to recommend.

4.2.2 Collaborative Filtering (CF)

In Recommendation system the data of the information to be recommended and the data of users need a recommendation are the known problems. The lack of information of news user and new item needed by users are also the common challenges in the recommendation system. The most popular approaches in this system is CF and it is modified by many researchers to overcome these usual problems.

CF is one of the most commonly used methods in personalized recommendation systems to suggest users by considering the assumption of the users with similar history of rating to have similar taste in the future. This means, CF algorithm recommend items based upon estimations of history of people with similar tastes. Recommender systems need to store certain information about the user preferences, known as the user profile to achieve this personalization [6]. In the proposed news recommendation system the collaborative Filtering approach considers the assumption of that users who rated or read news articles have the same taste of news for the upcoming news.

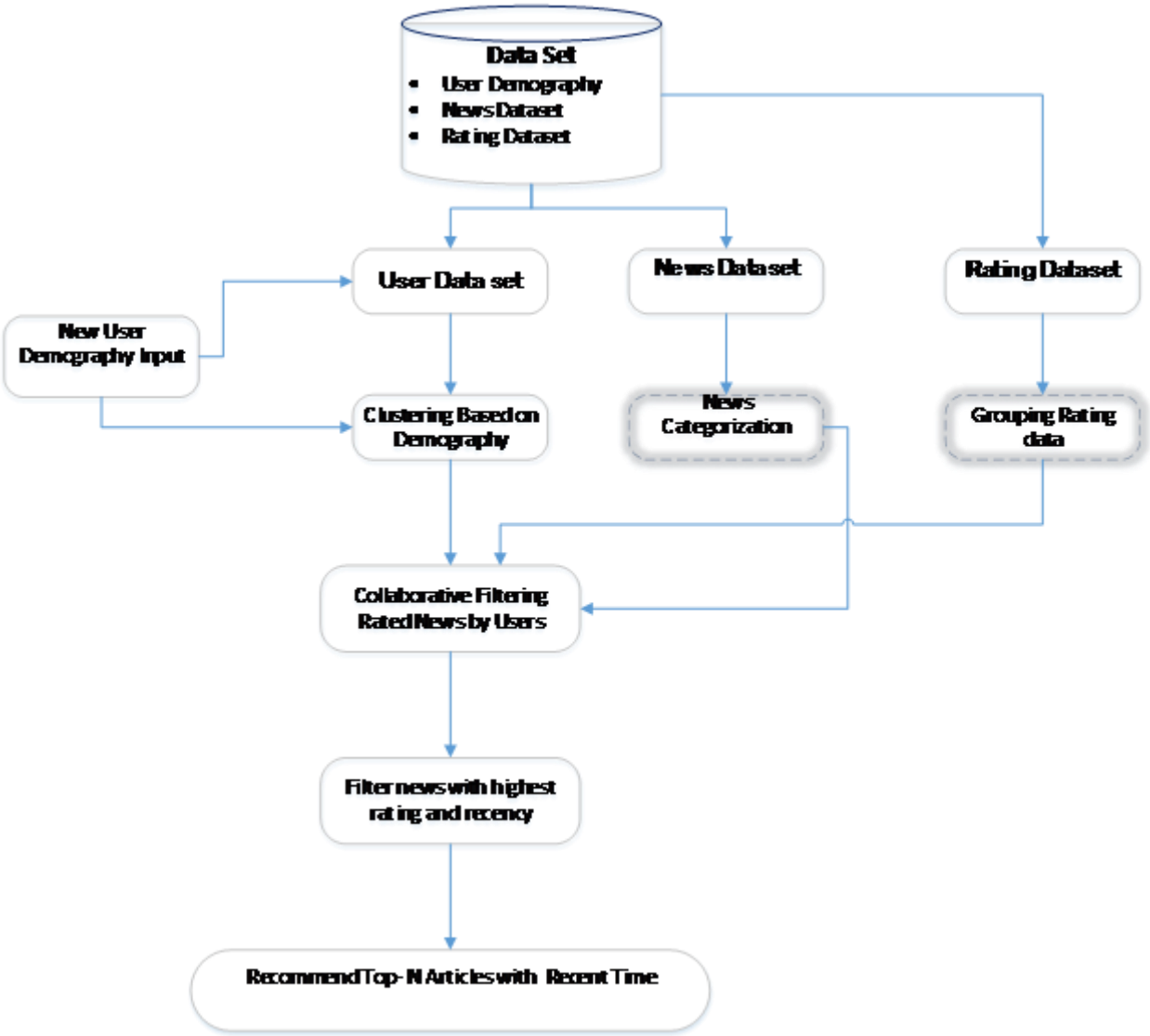

Figure 4.2 collaborative Filtering

In case of new user cold start, the problem happens when new user's for the system is introduced and there is lack of background information about the user's including the items/products he/she is searching for. In our case, the new user is the problem occurs with the lack of rating data for the news articles in the system. So, the provided recommendation couldn't be the personalized recommendation which means, the recommendation based on user's interest.

The collaborative Filtering needs user similarity for predicting the news to be recommended based on the news rated by similar users. To find out the similar users firstly, we need to have the users demographic since our problem targets recommendation for the user who had no any history in the system. In most currently used systems, demographic

information like user age, profession, location and sex are used as users' attributes to calculate similarity among users.

We applied the CF approach in our works to overcome the problem of new user cold-start. We consider the similarity of the users by their past history of the news they rated in their past time which is stored in the rating database. So, the new users who have no history in the system should be added to the active users. Active users mean, the users those have the rated news and the history information in the system. To add the new users, we need to have information used to add to the active users based on the similarity. Since we have the assumption of the users with similar demographic information will have the similar taste of news articles, we considered the demographic information as the information used to measure the similarity between the active and the new users. So, their similarity is done by their demographic information clustering. The clustering of the user's dataset based on their demographic information is done as it is described in section 4.4.1.

Then, after the new user added to the system cluster is identified the news rated in this cluster is filtered and the filtered news is predicted to the new user. But the generation of new for this users should consider the time of the news published and the rating value given by the active users. This approach need both the online process to automatically cluster the new user into the active users and to the offline process fetch news rated and stored on database. In our proposed work, we applied both memory-based and model based algorithms. The memory-based is used while our system filter the news from the memory and whereas, the model-based is applied while clustering the users in online by learning the user's neighbours from the system.

A. Memory-based CF

This technique uses the data stored on the memory from the database and process them to recommend the item for the user considered to this rated item. In our case, we applied this for the rated news articles which are stored on the database and we predict this news articles for the more similar users based on the history of rating the users had in the database.

B. Model-based CF

The other methods of CF in recommendation system which learns the user interest while processing to predict users interest or the news to recommended for the users. The users clustering process in our proposed system done by this model-based since the system learns the user appropriate groups in online after accepting user demographic information from the user system.

In our proposed algorithm this model-based CF needs another history of users since the new users has no any history in the system. We added the user demographic information in addition to rating data of existing users. The users clustered based on their demographic data and the existing clustered users have the rating value. So, it is possible to predict rating values of new user based on rating values of existing users.

C. Hybrid of Collaborative Filtering

This the approach of Collaborative Filtering which combines both the techniques of the model-based to cluster the users in online and memory-based to recommend news from memory.

4.2.3 Hybrid approach

As its name suggests hybrid technique is the value of different techniques combined together to have the better of 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 collaborative Filtering identifies similar users depending on the news they rated and content based Filtering identifies the users who rated similar news categories.

4.3 Proposed System Architecture

The proposed system architecture presents the model and the algorithm used to achieve the task to be done in the work. Since the problem is to solve the new user recommendation the work includes many components.

As we discussed in Section 2.5 the cold start problem is the common problem in recommendation system, since the recommendation needs data of the product to be

recommended for the user and the user data by itself to provide items back. 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 recommendation system is the interactions of the user with the product or information to be recommended. In our case, most news readers need the only news which is more interested for them. When the readers are new for the system, they face the cold start problem because of the lack of preference information. So feeding the users information is one way to overcome this problem using the user demographic information.

The entire proposed system architecture is shown as on Figure 4.3 and each of the main components are discussed in the next section.

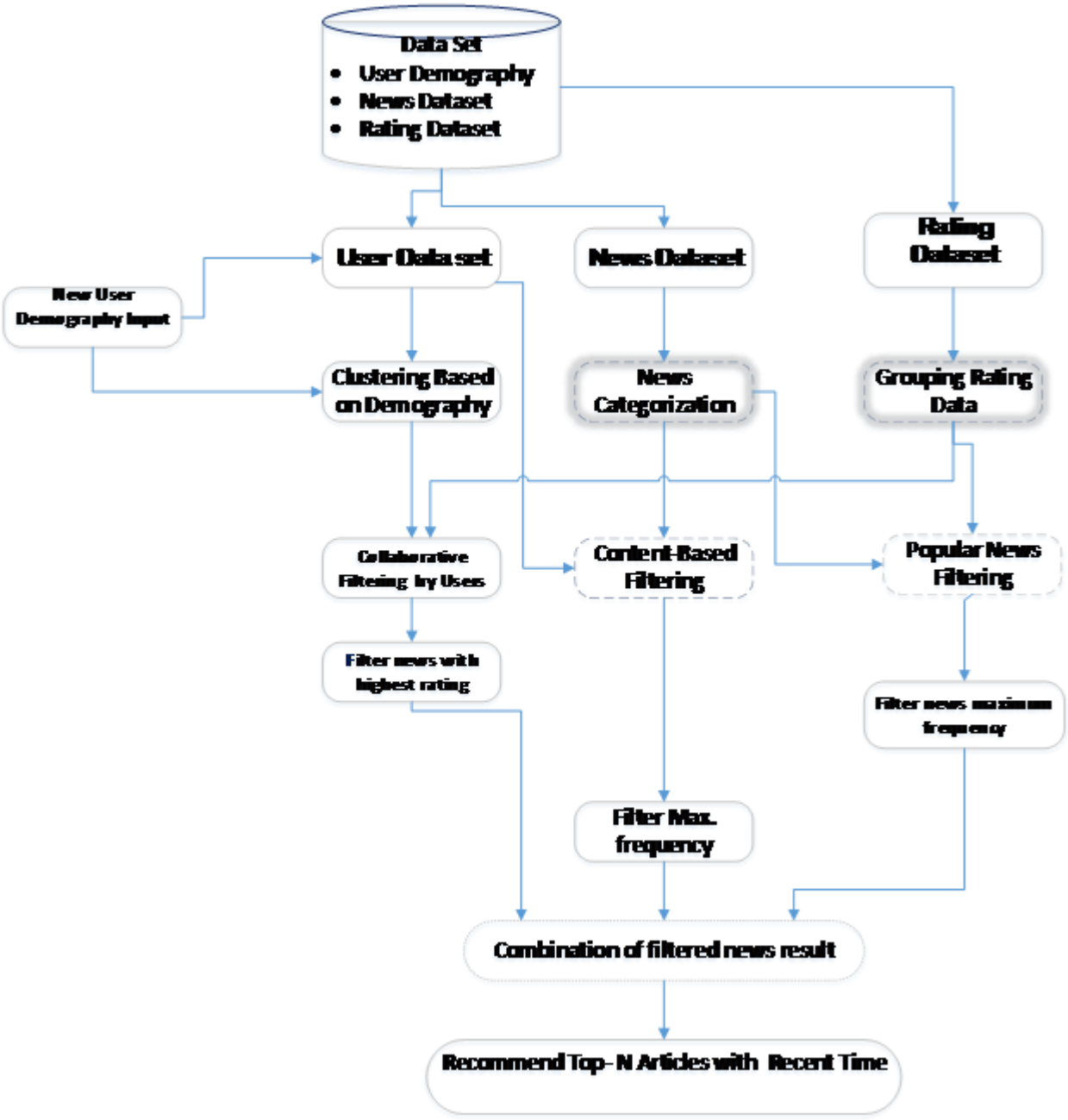

Figure 4.3 Proposed system Architecture

4.4 Components of Proposed System

The pre-processing of the dataset, accepting the new user demographic information and identifying the user cluster and finally recommending the interesting news articles for the users are the major tasks to be done in our proposed system. They are discussed in detail in next section, why they are needed and how they are functional by supporting them with algorithm.

4.4.1 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. We used the user demographic data, news articles and rating data of news articles by users. The number of users we used in this data is not easy to process and it is difficult to search compare the similarity of each user with all existing users. So, we applied the clustering algorithm to find the similar users based on their demographic data. We used clustering algorithm to pre-process the existing user dataset by using the similarity of user demographic of data in the database. For the similarity we selected the appropriate attributes used to cluster the data. 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 third and the last data set we used for our work is the rating data achieved by the interaction of news and the readers. And we group them based on the rate values.

a. **Clustering Users**

We clustered existing user dataset by grouping users based on the attributes we selected from user demographic information to cluster them. The existing users are clustered to their appropriate group based on their similarity according to their occupation, sex and age group.

We clustered users based on their professions, sex and age. We select these attributes since we have the assumption of the person with similar professions have similar interest of news because people want to get information about their professions every time. In addition to this profession data we cluster users according to their sex since male and female have different interest of news for their personal life. For example, female users want news of fashion for female and many news related to female users. In addition to this, age also has influence on user interest as young users and the old ones have different interest of news.

This clustering is important and we use it when we search similar users for the new registered users to reduce search complexity to find the specific user groups needed to be recommended. Since we are recommending new user, it is better to know the user's group by identifying the registered information about users.

We always check the similar groups for the registered new users based on profession, age group and their sex on online process. But the existing users are clustered to their appropriate group based on their similarity according to their occupation, age group and sex offline before the new user registration process. When new user registers the process of searching for the appropriate cluster of user will continue and if cluster is found the recommendation will be processed using the history of existing user similar to the registered new users. But if the new user couldn't get the exact group, a new groups for the users is created in our system. In addition, the recommendation provided for this new user's is the popular news filtered. The algorithm for user clustering is as follows.

Algorithm 1 *Pseudo code of User Clustering by Divisive Hierarchical Algorithm*

```
Given A set of user Profession P and Set of user Sex S
For each new_user Un
      IF profession Pn exists in P, THEN
            FOR each P Pi up to Px
            IF Un S==Male, THEN
                  Store Un data on Pi _Male
            ELSE
                  Store Un data on Pi _Female
                  REPEAT until i==x
            END FOR
            END IF
      ELSE //or if profession Pn not exists in P
      //Create new Cluster Pn
            IF Un S==Male, THEN
                  Store Un data on Pn _Male
            ELSE
                  Store Un data into Pn _Female
            END IF
      END IF
END FOR
```

b. Categorizing News Data set

The news dataset contains many news articles read and rated by different users. So, to access the news currently existing in the system we need to have a better way to process and retrieve from the dataset. This problem follows the scalability problem in news recommendation. Since there are many articles and takes too much time and memory for processing and retrieving them. The better way to overcome this problem is using grouping or categorizing the news article data based on their categories.

The news articles in the dataset contains different features like time, categories, title and rate value. Therefore, we can group the dataset based on the categories. We grouped this news dataset based on the category of news article. This means, the news with similar categories are grouped under the same group. For example, the news articles with business issues should be considered under business group and the same for sports and others.

Generally, the 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 news recommendation system.

c. **Grouping Rating Data**

The other data set we need to cluster is the rating data of news rated by users. This data set contains the news rated by users with the user ID and news ID or news URL. In addition to this, the rating data contains other features like time, rate value and news categories. Since it contains the data of many user's ID rated news and many news rated by many users it is the huge data.

In the proposed system, the new user registered for the system looks for news and the system provide news requested by user based on their rate values. So, the system need to process the rating data which is huge. The processing of this huge data takes many time and memory which follows the scalability problem. Therefore, it is important issue we need to consider in the study and we used grouping data method for this dataset. We apply method to group the rating dataset.

The grouping data methods for this dataset is done based on the rate value of news given by the users who read and rated the news. Since, the grouping considers the values of the news rating given by the users and we need news with more rating values this also reduce the news those have less rating values in our works. After we grouped them we have 4 groups of news. Those are above average news rating data which contains news rated by user with the 4 and 5 rating value, average rating data with 3 rating value, below rating data with 1 and 2 rating value and visited rating data with no rating value but visited by the readers.

4.4.2 Registering New User

We focused on user data to achieve our main objectives of this study. In this study our proposed system mainly used with the user side information. The user information we used for our study is the demographic information data. The user we need to recommend in this study is the new user who have no information from the system provides the recommendation. Registering user enables us to collect the new user demographic information.

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

4.4.3 Clustering New Users

The registered user groups should be identified according to user similarity and the data are registered to the appropriate cluster or the similar groups. This identification of new users is done based on demographic attributes. If the registered new users most attributes are similar with existing user's demographic data, the users will belong to that clustered existing users. Then the system adds the user demographic data into his/her respective group.

The users clustered under similar users based on their professions and sex as explained in section 4.4.1 may be many users and we need to reduce the number of users under the selected clusters. Filtering of this information will reduces the number of users by assuming that users with similar age groups will have similar interest of news articles to be read. Firstly, we grouped users into four age groups. The first group with the age range of 15-30, the second group with the range of 31-45, the third one with range of 46-60 and the last group with 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 are fetched from the rating database. This clustering is done online for new users.

4.4.4 Retrieving Rated News

Since the recommendation should have to provide the news, retrieval of news should be done by the system using the proposed model. So that the filtered users based on age group from the particular cluster get the rated news. The rated news by filtered users should be fetched and the process of priority consideration is applied on them before predicting to the new registered user. Since the news articles may not be recent and may not have the highest rating value, we need to identify according to their publish time and their rating value they have got from users. Retrieving news is done in both content-based and collaborative filtering approaches.

I. **Filtering news by content-based approach**

The news categories used to cluster the news rated by users and this news are rated by different users. So the news rated by many users from one categories should be recommended for the new users similar in the cluster in addition to recommended based on the collaborative filtering.

II. **Filtering news by collaborative filtering approach**

News rated by the user under one group according to their demographic similarity are filtered. Since the rated news are clustered into four different groups, the system should check from all groups and follow the priority to return the news. The news from highest rating value should get the first priority if it fulfils the number of news to be generated for the readers. If not the next group is checked and also it should 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 the news since users need the recent news. So, the recent news which have highest rating value are generated for the readers.

III. Filtering popular news

The popular news assumption in this work is the news articles with highest rating value and rated by many users in recent time. 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 by many users and it is popular with many readers. Finally, we

retrieve these articles by checking their rating value and retrieve all with highest rating value. The following pseudo code shows the algorithm we developed for recommending popular news.

Algorithm 2 *Pseudo code of Popular News Recommendation Algorithm*

```
Input: set of user User-News Rating cluster Rc and set of News category
CN
// recommend popular news
FOR each news category CN
      Find news with highest frequency PHF
      FOR each PHF
            Find news with highest rate value PHR
            Store on popular news PN
      END FOR
END FOR
      Display PN by time descending order
```

4.4.5 Combining Results

To overcome the separated approach of news recommendations we applied the hybrid approach. The hybrid approach is done by combining the different results obtained by the separated approaches we used in our study. The combination of the result of both news recommended based on rating prediction and the 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 both predicting and popularity to be recommended after ranking by the time value order.

4.4.6 Generating Top Recommended News articles

It is the stages to provide the news titles which are selected to be recommended by the system for the readers is done by top values method. The top N articles recommendation is done by the ranking algorithm we developed for news recommendation system. Since the information our work recommend is the news we need to consider the time the news articles published and the popularity of the articles by readers.

The news articles to be recommended is retrieved through both the two approaches content-based and collaborative filtering and the news articles generated based on

popularity. Since each of the approaches we have used have its own ranking methods for the articles to be generated and the results obtained are generated by the ranking algorithm of each of the approaches used. And the results obtained in each individual approaches are combined together by aggregating their results. Finally, we generate the news articles to be recommended by their time they were published. This means the recent news are displayed at the top. Finally, the system stores the recommended on log files of the system to reduce the search complexity for next search.

4.5 Algorithm of the proposed model

The algorithm of our proposed system which consider all the approach we used to overcome the problem we 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 cluster Rc, set of new User Demographic
data Un, set of news category CN and User Cluster UC
OUTPUT: set of news articles
BEGIN
FOR each new user Un
      FOR each of user cluster UC
      Find new user cluster Unc
      IF (user Unc not exist) //if new user has no similar User from
      existing user cluster
      FOR each news category CN
            Find news with highest frequency PHF
            FOR each PHF
                  Find news with highest rate value PHR
                  Store value on PN
            END FOR
      END FOR
      Display PN by time descending order
      END IF
      Else //if user has same cluster from existing users
      //Filter news based on user age
      IF (user Un .age<=45) // user is young
            FOR each News category CN
                  Find news with highest rate value PHF
                  FOR each PHF
                        Find news with highest frequency PHR
                        Store on PN
                  END FOR
            `END FOR
      //find news based on news category to find content-based CB
            FOR each News category CN
                  Count news articles
```

```
        END FOR
              Find news category with highest frequency C_HF
              Store on CB_N
    //find news based on user cluster to find collaborative filtering
    CF News
              FOR each rating cluster 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
    //find news based on news category to find content-based CB
        FOR each News category C_N
              Count news articles
        END FOR
              Find news category with highest frequency C_HF
              Store on CB_N
    //find news based on user cluster to find collaborative filtering
    CF News
              FOR each rating cluster 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
```

4.6 Summary

We have developed new model used to recommend news to new user. We have used the dataset of the existing user and news with the rated news to cluster the users and news according to their similarity and finally we recommend by predicting the new user based on similar user of the demographic information with existing users. We identified each of the components of our model and with architecture we proposed and we put the entire algorithm of the proposed.

The scalability problem in the news recommendation is fixed by using clustering methods. It is done by clustering the dataset used in this study.

The user should register to get recommendation system by submitting the demographic information into the system through the user interface. Then, the system clusters the new

user registered into the appropriate cluster based the demographic information submitted by user.

The users are filtered based on the registered user age group from the clusters of the user and the news rated by the filtered user will be retrieved and these news articles are predicted to the new users by collaborative filtering approaches. The news clustered based on the news category and the category with many users are predicted to the user by content-based Filtering.

Popular news are retrieved based on the frequency of news rated by users and the rate value. By combining of the two approaches result together and displaying them in descending order of the time they are published on the web sites.

Chapter Five

Experimentation Result, Discussion and Evaluation

5.1 Overview

This chapter present the implemented model and experiment in this study with the result obtained for evaluating the performance of the study regarding to the problems we studied and comparing with other previous works done by other researchers. To do so, we need to have analysis of our data set that we used for implementing our model. The result generated by the implementation are discussed and the evaluation of the result obtained are done by using the metrics we used according to our problem dimensions.

5.2 Experimentation

In the next section we discuss the dataset used for the experimentation, the tools used for implementation and the evaluation metrics used to measure the performance of the proposed system.

5.2.1 Implementation Tools

We have used different tools to implement the prototype for both backend and frontend. The implementation tools we used for our study is the Java Programming language with *NetBeans 8.0.1* tools for front end to develop the model and the *MySQL 3.1.0* for backend as the database to store the dataset and to process them by connecting with NetBeans. NetBeans tool is used to implement the prototype of our model and to register new user and provide the output for users.

5.2.2 Data Set

To overcome the new user problem, we have prepared three types of data which are news dataset, user demographic information dataset and rating dataset.

The data source for our system is the dataset found from popular GitHub websites which collects data from different websites and search engines, and provides news dataset from Seven Search Engine websites. This dataset contains the text file of news with the news categories, date published, source of news, URL of news articles, Description text of news

articles and the titles of articles. We have changed the text data of news from the source into table in our database according to their attributes appropriately.

In addition to this, we have demographic information of users those who rated or read the news with their sex, age, profession and user ID. The demographic information of users stored in another table. The rating value given for each news articles by the active users are also another data we used in this study and it is stored on another table. This rating data contains the user ID who rated news articles, the rated news article URL, the Date of publish and the rate value attributes.

a) **User demographics dataset**

It is data of users which includes user age, user location, user ID, user occupation and sex. It is important in our works since our work purposely based on the user problem and we have to have their demographic information. As we mentioned, our objective is to solve the new user problem which happens by the lack of information in the system. So, we need to have personal information to predict what will be their interest news to be read based on the similarity of them with existing users according to their personal or demographic information. Since, we assumed that some users may have similar interest with their age groups or their similar profession staff. The dataset contains four attributes; those are User ID, Age, Occupation and Sex. They are described by their function in our system and their data types as follows.

i. User ID: - It is the primary key in the table and used to identify each of users in the system and it is given by the system. Its data type is character. It is important to know the user who rated the news article in the rating data using user ID as a foreign key.
ii. Profession: - This is the information which is used to identify the profession or the jobs of the user to suggest the news related to his/her occupations. The users in the dataset are clustered based on this attributes and, why it is selected will be discussed in the Attribute analysis section. It is a text data type.
iii. Age: - For identifying the interest of readers in similar age group since, users in similar group may have similar interest in our assumption and it is also analysed in next section, why it is used to cluster the users. It is the integer value.

iv. Sex: - The gender of the user is also another data that will influence the interest of news categories since female and male interest is different. This will be analysed in next section with reasoning.

Generally, user demographic information is the information used for finding similar users and clusters them according to their demographic information similarity for this work.

b) News dataset

The news articles data collected from the Seven Search Engine Website with their URL, Category of news, Date published, their sources and their titles. Let us describe each of their attributes and their function in our system.

i. URL (Uniform Resource Locator): - It is the address of the news web page for detail information about the news articles. It is used as a *primary key* for our system in the news dataset and as a *foreign key* in rating dataset. Its data type is character.
ii. Category: - In our system, identifying the categories of the news is one of important task since it is used to know which category is more interested by the specified user groups. The categories of news in our data are Sport, Health, World, US (United State), Business, Science and Technology and Politics. Its data type is text.
iii. Date Published: - It is the attributes which is used to determine the freshness of the news since our system gives priority for the more recent news. It is data type is timestamp and contains (Year, Month, Day, Hour, Minutes and Seconds it is published).
iv. Titles: - is the title or topic of the news articles and it is the returned values to be displayed. In addition, it is used to extract the most keyword in the last time read news for the purpose of recommending the popular news. Its data type is text.

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 purpose and for reading or watching videos. In our case, the rating data is the dataset which holds the value of rating given by users for specific news article by existing users for existing news articles in the dataset. 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.

i. User ID: - The attributes used to identify each of users in the system and it is given by the system. Its data type is character. It is important to know the user who rated the news article as a foreign key.
ii. URL: - It is the address of the news web page for detail information about the news articles which is used as a *foreign key* in in this data set to identify the rated articles by the users. Its data type is character.
iii. Rating value: - The value that is given by the user for each of news articles and it is the integer value in between 0 and 5. It is the attributes used to cluster the rating data set to put the news that have relate value in similar groups.
iv. Category: - The categories of new rated will be identified by this attributes since our system need the average rating value of similar categories in similar user groups. It is text data types forwarded from news dataset.
v. Date: - The date of the news published to identify the freshness of the rated news. And its data type is timestamp and it is the attributes forwarded from news dataset.

This dataset is divided into two parts; one for training dataset and other for testing dataset by removing the user rating values from rating dataset to consider the user as a new user.

5.2.3 Evaluation Metrics

To evaluate the performance, we depend on the objectives of our study and we selected the related metrics to our objectives. Since the main objectives in our study is to recommend the more related or interested 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 the precision,

recall and F-score. The following figure shows the descriptions of all metrics with their relationships [54].

Table 5.1 Evaluation Metrics

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

A. Precision

Precision is the most performance measurement in both traditional information retrieval and the different recommendation system and which measures the performance of the system that provides the information with the relatedness of the information retrieved. So, in our case the news articles recommended is measured how many of news articles are correctly recommended out of the all recommended articles based on the new user interest. This result will be find by calculating using Equation 1, [54].

$$\boldsymbol{Precision} = \frac{\boldsymbol{Good\ Articles\ Recommended}}{\boldsymbol{All\ Articles\ Recommended}} \ldots \quad \ldots. \text{(1)}$$

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

B. Recall

Recall is another measurement of Recommendations to measure the performance of many system and it measures how much items are correctly retrieved in information retrieving system or recommendation system. In our case, it measures the good recommended out of all good recommended items as described in the Equation 2, [54].

$$\boldsymbol{Recall} = \frac{\boldsymbol{Good\ Articles\ Recommended}}{\boldsymbol{All\ Good\ Articles}} \ldots \quad \ldots.. \text{(2)}$$

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

C. F-Score

F1-Score is the harmonic mean value of both precision and recall results as it is formula is shown in Equation 3, [54].

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

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

5.3 Experimentation Result

The results we obtained in our study is analysed as follows to show what achieved and how much the problems stated are solved by our model. The discussion and descriptions of the result obtained will be explained more in figure or table format with more explanations. The performance results we obtained in this study according to our metrics is discussed.

We used 160 active users which contains different professions, both sex and different age groups. These selected users are the users which rated 32,624 news articles and make 303,594 rating data. The users clustered into 38 groups based on the 19 profession with each to both female and male sex. The news articles clustered into 7 groups based on the 7 news categories. The rating data clustered into 4 groups based on the rate value ranges. The rate value with 1 and 2 into below average, rate value with 3 into average group, rate value with 4 and 5 into above average group and rate value with 0 or news articles not rated by users into visited group.

The system generates the news for the new user registered and it clusters the user into the appropriate cluster and if the cluster of the new user is not available in the system it creates a new cluster and add the data to it. The clustering algorithm is done by our algorithm to cluster the users online after registration is done. When new user registered to the system and asks for recommendation the system needs the user 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.

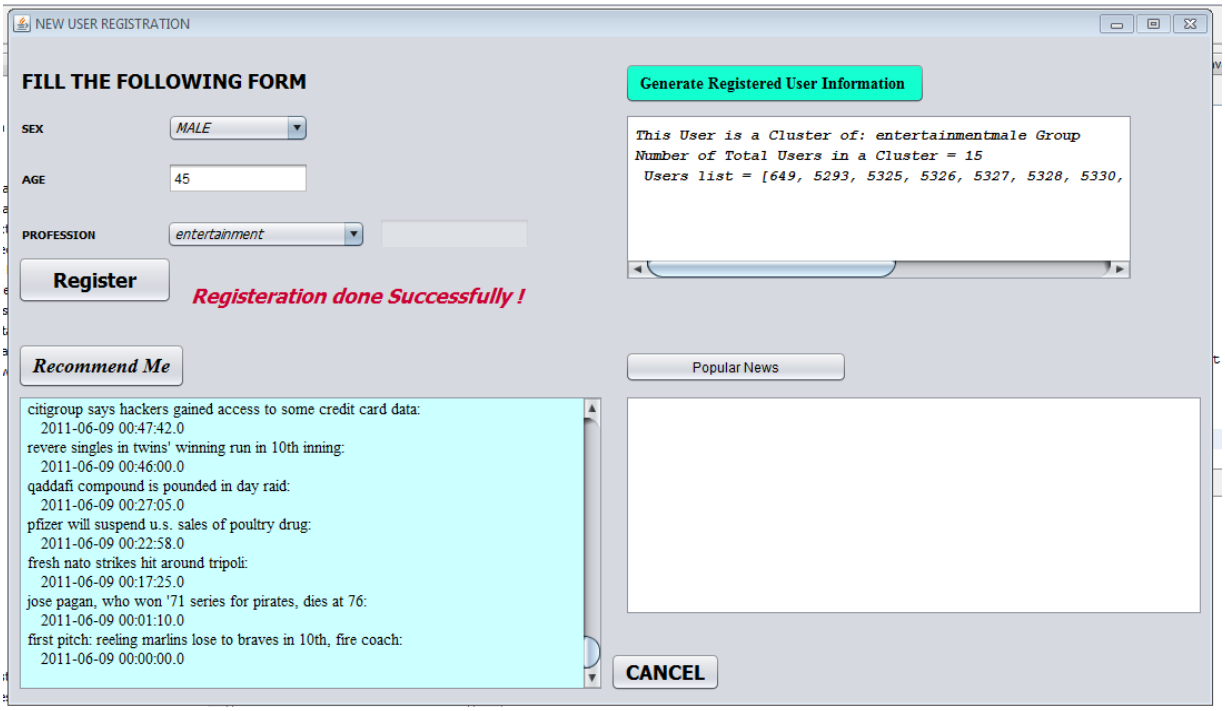

Figure 5.1 User Interface Screen shot

The model we developed in this study uses the data mentioned at above to provide the output as the above figures from screenshot for the users, and we evaluate the performance of final output result provided for users by the metrics we have used to measure the performance of the system. We evaluated the performance of our works in two ways. One is the experimentation of our work for comparing each users with each other and the other is the experimentation of the system for comparing the performance of the users with clusters which consists of many users.

5.3.1 Experimentation for individual user similarity

For the individual users we evaluated by taking 16 which means 10% of active user in dataset which contains 160 active users. For preparing testing dataset we used these 16 users rating data. 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 proposed model to recommend the news for the users and we compare the previous news rated by the user with this actual recommendations. Then, we calculate 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 shown as 5.1.

Table 5.2: 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 5.1.

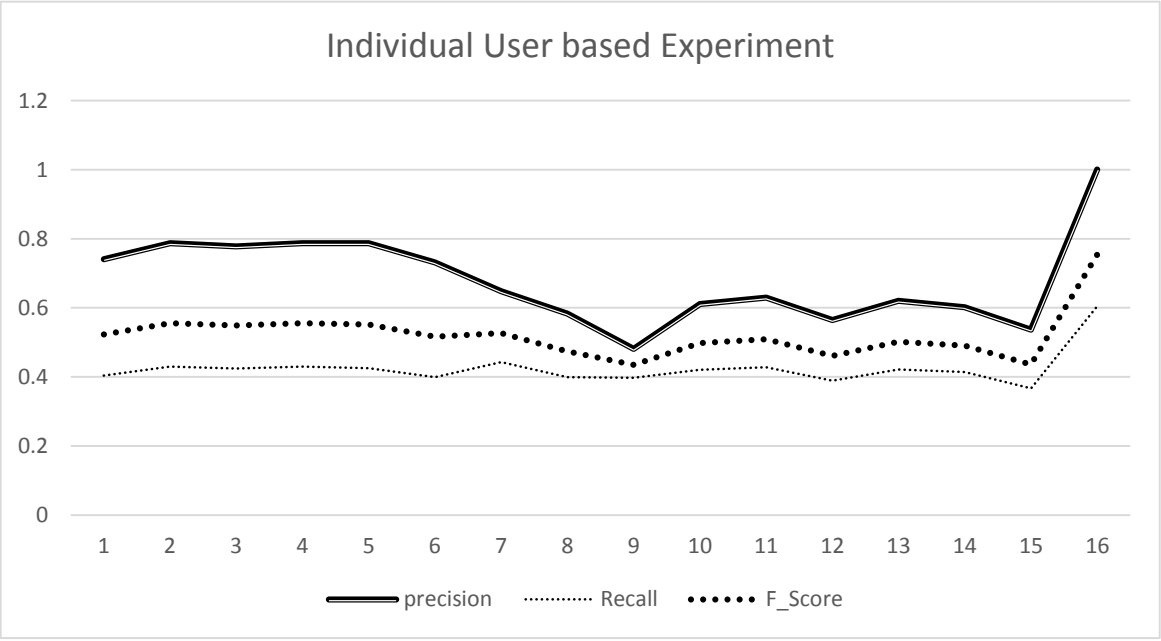

Figure 5.2 Experimentation value based on individual user similarity

5.3.2 Experimentation by user cluster based similarity

The other way we evaluated our performance is based on the user cluster. We took 5 user groups which of 10% of user groups out of user clusters we have in our dataset. Then, we recommend some of new users similar to the cluster selected and we compare the accuracy performance by comparing the actual recommendation with the recommended for that cluster. According to this experimentation, results of the work is explained is shown in Table 5.2.

Table 5.3: Experimentation values for cluster based user similarity

Cluster 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 5.3.

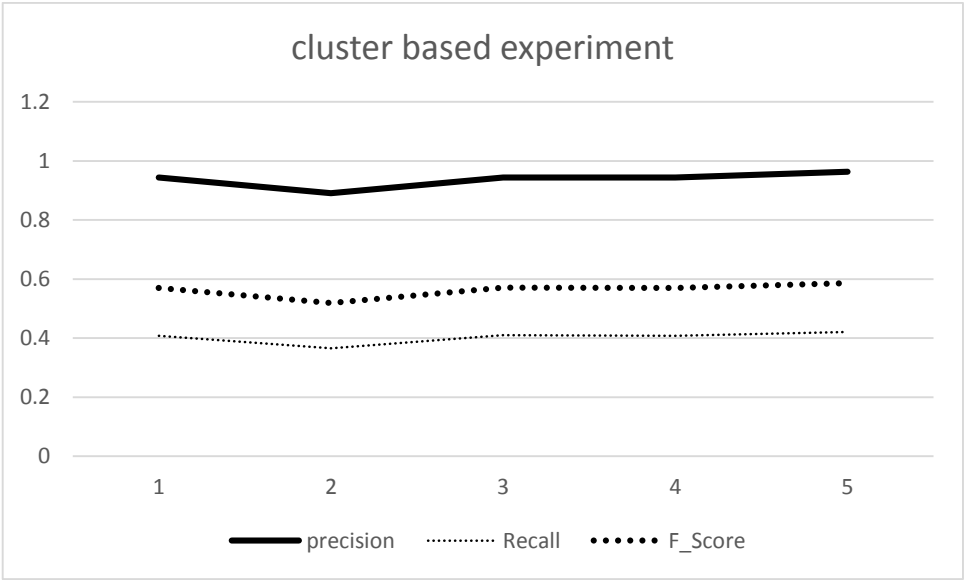

Figure 5.3 Experimentation value based on user cluster similarity

5.4 Discussion

The result of our experiment in this work contains two experimentation way and each has different values as discussed in previous section. 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 clustering based and individual user similarity.

According to the two experiments results, we analysed that the more accurate recommendation is done for the users in the clustering 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 cluster based recommendation contains more related news regarding to the users in that cluster more than the news recommended for individual user similarity. According to this value the Recall in the individual performs less performance than in cluster-based. The reason behind this result is, the finding of the good recommendation result numbers from many users in the cluster. So, if the good news articles recommended are many then, the Recall value will become less.

Comparing Result Performance

The previous study result performance done for recommending news and to solve new user problem are analysed as follows.

1. Hee-Geun Yoon [47], entitled Personalized News Recommendation using User Location and News Contents. They propose a novel model to incorporate user location into a user preference for the location-based personalized news recommendation. In the evaluation of their model, it is shown that combined model outperforms both STPM and LDA. These experimental results performs for all k's in both locations except k = 1 at home. Its average NDCG@k is 0.69 at office and 0.74 at home.
2. S. K. Tiwari and S. K. Shrivastava [49], are studied an approach for Recommender System by using Collaborative filtering techniques which using user demographics and items genres to solve new user and new item. MovieLens dataset which contains three sets those are rating data, user demographic and item data was used

for their evaluation. The system is evaluated using Mean Absolute Error (MAE) and performs 0.7963.

3. Darvishy *et al.* [52], studied on New Attributes for Neighbourhood-based Collaborative Filtering in News Recommendation. The authors utilized new attributes in addition to standard recency and popularity such as Reading Rate and Hotness. This is carried out via experiments using k-means. Then they had performed the 0.12654 average value for precision, 0.20674 average for Recall and average for 0.15109 F-Score.
4. Sunitha and Adilakshmi [57], proposed a new approach to use user's side information in addition to user-item rating matrix to address new user cold-start problem. User's side information is obtained from Social Networks. Results showed that both personality-based similarity and the hybrid scheme performs precision of 0.08149 and Recall of 0.16655.

The following table shows the result performed by using different techniques and parameters to recommend news and solve new user problem.

Table 5.4 Result comparison

Authors	Performance metrics				
	precision	Recall	F-Score	MSE	NDCG@K
Hee-Geun Yoon [47]	-	-	-	-	0.69, 0.74
Darvishy *et al.* [52]	0.12654	0.20674	0.15109	-	-
S. K. Tiwari and S. K. Shrivastava [49]	-	-	-	0.7963	-
M. Sunitha and Dr. T. Adilakshmi [57]	0.081498	0.16655	-	-	-
Proposed hybrid with demographic data	**0.6805**	**40.246**	**0.521**	**-**	**-**

Our study is different with the objectives and we evaluated based on our objectives. Since we have begun the study by different objectives the data set we used is not used in other study of recommendation. Generally, the performance of our study performs the good recommendation accuracy.

5.5 Summary

We have discussed the dataset, tools we used for implementing the model stated and we discussed the performance evaluation of the results obtained. The results obtained are also discussed and showed on screen shot figures and table formats. The main concepts discussed in this chapter are about the result analysis using different metrics to evaluate the performance of the proposed model as summarized below.

The output of our work with the user interface designed is discussed and the performance of the work is also evaluated and explained in different tables and figures forms by discussing the metrics we used according to our objectives. Each of the metrics used are also discussed with their appropriate formulas.

In addition, result analysis, the dataset used and the obtained output by the algorithm proposed is discussed with the screen shot figures. The evaluation metrics used are discussed and the experimental result of them is also identified for all the way the system evaluated. So, the performance of this study find the Precision value 68.05%, Recall 42.46% and 52.1% of average of F1-Score for the user similarity based on individual users in the system. And for the second way of our evaluation which uses the similarity of users based on the similarity of users within the same cluster and the value obtained is precision of 93.75%, average recall of 40.25% and average F1-Score of 56.31%. So, the recommendation provided based on the user cluster is better than provided by that of individual user similarity.

Chapter Six

Conclusion and Future Work

6.1 Conclusion

In this thesis, User Demographic Data with Hybrid news Recommendation system is proposed. This hybrid recommender scheme combines content-based and collaborative approach with user demographic information. The content-based component uses the news features category to get knowledge about the content type of the news to select for the recommendation. The content-based uses the similar user in the same cluster 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 neighbour users and/ or similar users from existing users by using clustering algorithm to cluster the user neighbours. 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 using clustering algorithm included in this work to cluster 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 results 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 results. Moreover, the proposed model performed Precision value of 68.05%, Recall value of 42.46% and F-Sore of 52.1% for individual user similarity experiments. And the user cluster based experiment

performed precision value of 93.75%, recall value of 40.25% and F-Score value of 56.31%.

In this proposed model, we include User Demographic Data filtering approach towards recommending relevant news articles for the new users, in addition to combining content and collaborative filtering approach to form a hybrid approach. 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.

6.2 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.

Reference

[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] 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.

[5] L. Sharma and A. Gera, "A Survey of Recommendation System: Research," *International Journal of Engineering Trends and Technology*, vol. 4, pp. 1989–1992, 2013.

[6] J. A. Gulla and R. C. Erdur, "A Survey on Challenges and Methods in News Recommendation," *in: Proceedings of the 10th International Conference on Web Information Systems and Technologies*, Vol. 2, pp. 278-285, 2014.

[7] Resnick, P. and Varian, H. R., "Recommender Systems," *Communication ACM*, vol. 40, pg. 56-58, 1997.

[8] Ahn, H. J., "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem," *Information Sciences*, pp. 37-51, 2008.

[9] P. Melville and V. Sindhwani, "Recommender systems," i*n Encyclopedia of Machine Learning*, *Springer*, pp. 829-838, 2010.

[10] Kantor, P. B., Rokach, L., Ricci, F., & Shapira, B., "Recommender systems handbook Springer," *Database Management & Information Retrieval*, ISBN 978-1-4899-7637-6, 2011.

[11] 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.

[12] https://en.wikipedia.org/wiki/News.

[13] https://en.wikipedia.org/wiki/Online_newspaper.

[14] Mansi Sood, Harmeet Kaur, "Survey on News Recommendation," *International Journal of Advanced Research in Electrical Electronics and Instrumentation Engineering*, Vol. 3, Issue 6, 2014.

[15] 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.

[16] Le Hoang Son, "Dealing with the new user cold-start problem in recommender systems: A comparative review," *in: press. Information Systems*, 2014.

[17] 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.

[18] K. G. Saranya and G. S. Sadhasivam, "A personalized online news recommendation system," *International Journal of Computer Applications*, pp 6–14, 2012.

[19] M. Balabanović and Y. Shoham, "Fab: Content-based, collaborative recommendation," *Communications of the ACM*, vol. 40, pp. 66-72, 1997.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] D. Maltz and K. Ehrlich, "Pointing the way: Active collaborative filtering," *in Proc. the SIGCHI Conference on Human Factors in Computing Systems*, pp. 202-209, 1995.

[26] Shardanand, U. and Maes, P. Social information filtering: Algorithms for automating "word of mouth," *in: Proceedings of ACM Conference on Human Factors and Computing Systems,* pp. 210–217, 1995.

[27] Herlocker, J. L., Konstan, J. A., Borchers, A., andRiedl, J. "An algorithmic framework for performing collaborative filtering," i*n: Proc. of SIGIR,* 1999.

[28] Linden, G., Smith, B., and York, J., "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet Computing,* pp. 76-80, 2003.

[29] Hofmann, T., "Latent semantic models for collaborative filtering," *ACM Trans. Info. System,* vol. 22(1):89-115, 2004

[30] P. Cotter and B. Smyth, "Ptv: Intelligent personalized tv guides," *in: AAAI/IAAI,* pp. 957-964, 2000.

[31] 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.

[32] 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.

[33] M. J. Pazzani, "A framework for collaborative, content-based and demographic filtering," *Artificial Intelligence Review*, vol. 13, pp. 393-408, 1999.

[34] M. Balabanović and Y. Shoham, "Fab: Content-based, collaborative recommendation," *Communications of the ACM*, vol. 40, pp. 66-72, 1997.

[35] Y. Xue, C. Zhang, C. Zhou, X. Lin and Q. Li, "An effective news recommendation in social media based on users' preference," *in: Proc. Education Technology and Training International Workshop on Geoscience and Remote Sensing. ETT and GRS*, pp. 627-631, 2008.

[36] Urszula Kużelewska "Clustering Algorithms in Hybrid Recommender System on MovieLens Data," *Studies in Logic, Grammar and Rhetoric*, vol. 37, no. 50, pp. 125–139, 2014.

[37] D. P. Krishna, A. Senguttuvan, and T. S. Latha, "Clustering on Large Numeric Data Sets Using Hierarchical Approach: Birch," *Global Journal of Computer Science and Technology Software & Data Engineering*, vol. 12, no. 12, 2012.

[38] L. Rokach and O. Maimon, "clustering methods," *Data mining and knowledge discovery handbook*, ISBN. 978-0-387-24435, pp. 321-35, 2005.

[39] Fraley C. and Raftery A.E., "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis," *Technical Repor*t, No. 329, 1998.

[40] Han, J. and Kamber, M., "Data Mining Concepts and Techniques," *Data Mining*, ISBN: 978-0-12-381479-1, 2001.

[41] 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.

[42] 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.

[43] Salter, J. and Antonopoulos, N., "Cinema Screen recommender agent: combining collaborative and content based filtering," *IEEE Intelligent Systems*, vol. 21, pp. 35-41, 2006.

[44] 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.

[45] Pazzani, M., "A framework for Collaborative, Content Based and Demographic Filtering," *Artificial Intelligence Review*, pp. 393-408, 1999.

[46] 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.

[47] 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.

[48] Rong Hu and Pearl Pu., "Using Personality Information in Collaborative Filtering for New Users," *in: preceding of semantic scholar*, 2015.

[49] S. K. Tiwari and H. Potter, "An Approach for Recommender System by Combining Collaborative Filtering with User Demographics and Items Genres," *International Journal of Computer Applications*, vol. 128, no. 13, pp. 16–24, 2015.

[50] L. Safoury and A. Salah, “Exploiting User Demographic Attributes for Solving Cold-Start Problem in Recommender System,” *Lecture Notes on Software Engineering*, vol. 1, no. 3, pp. 1–5, 2013.

[51] S. Solanki and S. Batra,” Recommender System using Collaborative Filtering and Demographic Characteristics of Users,” *International Journal on Recent and Innovation Trends in Computing and Communication*, Vol.3, No.7, 2015.

[52] 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.

[53] S. Gong, “A Collaborative Filtering Recommendation Algorithm Based on User Clustering and Item Clustering,” *Journal of Software*, vol. 5, no. 7, pp. 745–752, 2010.

[54] M. Surendra and P. Babu, “An Implementation of the User-based Collaborative Filtering Algorithm,” *International Journal of Computer Science and Information Technologies*, vol. 2, no. 3, pp. 1283–1286, 2011.

[55] J. Bobadilla, F. Ortega, A. Hernando, and J. Bernal, “A collaborative filtering approach to mitigate the new user cold start problem,” *Knowledge-Based Systems*, vol. 26, pp. 225–238, 2012.

[56] H. Liu, Z. Hu, A. Mian, H. Tian, X. Zhu, “A new user similarity model to improve the accuracy of collaborative filtering,” *Knowledge-Based Systems*, vol. 56, pp. 156–166, 2014.

[57] 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.

[58] Le Hoang Son, “Dealing with the new user cold-start problem in recommender systems: A comparative review,” *in: press. Information Systems*, 2014.

[59] Wai-Ho Au, Keith C. C. Chan, Andrew K. C. Wong, and Yang Wang, "Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data," *Technical Report*, 2005.

[60] F. O. Isinkaye, Y. O. Folajimi, B. A. Ojokoh, "Recommendation systems: Principles, Methods and Evaluation," *Egyptian Informatics Journal*, vol. 16, pp. 261-273, 2015.

Declaration

This study is my original work and has not been submitted as a partial requirement for a degree in any University and that all sources of material used for the study have been duly acknowledged.

Declared by:

Zerihun Olana

Signature: _____

Date: _____

Confirmed by advisors:

Melita Luke (PhD)

Signature: _____

Date: _____