Web Personalization System for Online News

Hsu Myat Mon, Thin Thin Htike
Computer University, Pathein
eihsulay@gmail.com

Abstract

Web personalization can be defined as the process of customizing the content and structure of a Web site to the specific and individual needs of each user taking advantages of the user's navigational behavior. Recommender System applies knowledge of discovery technique to the problem of making personalized recommendation for information, products or services. The content-based filtering approach recommends the contents that the user likes in the past. Content-Based filter recommends items based solely on a profile built up by analyzing the content of items that a user has rated. The collaborative filtering approach recommends the contents that are liked by other users with similar interests. In this paper, an effective frame-work for combining content and collaborative filtering is used to predict news articles of interest for user. The proposed system is developed using the feature of web personalization to improve the recommend-dation process of the system.

1. Introduction

Recommender systems form a specific type of information filtering system techniques that attempt to recommend information items (music, books, news, images, web pages, etc) that are likely to be of interest to the user. A recommender system can be explained in term of the ways in which the item is similar to items the user has rated before [4]. Typically, a recommender system compares a user profile to some reference characteristics, and seeks to predict the 'rating' that a user would give to an item they had not yet considered. These characteristics may be from the information item (the content-based approach) or the user's social environment (the collaborative filtering approach).

Collaborative filtering is the technique of using peer opinions to predict the interest of others. User indicate their opinions in the form of ratings on various pieces of information, and the collaborative filter correlates the ratings with those of other users to determine how to make future predictions for the rater. The collaborative filter shares the ratings with other users so they can use them in making their own predictions. However, collaborative filtering cannot provide a prediction for an item when it first appears since there is no user rating on which to base the predictions. When CF system first begins every user suffers from the early rater problem or new user problem for every item [5].

Content-based filter recommends items based solely on a profile built up by analyzing the content of items that a user has rated. A content-based filter analyzes items rated by an individual user and uses the content of the items as well as the provided rating to build a profile to compare to other non-rated items to recommend additional items of interest. However, content-based filters provide recommendations merely based on user profiles .Therefore, user have no chance of exploring news items that are not similar to those items include in their profiles [1].

The advantages of the system, collaborative filtering systems can be enhanced by adding contend-based filters. By using a combination of content-based and collaborative filters we can realize the benefits of content-based filters which include early predictions that cover all items and users, while gaining the benefits accurate collaborative filtering predictions as the number of users and rating increases [3].

In this paper, Related works are discussed in section 2. Web personalization system is explained in section 3. Content-based filtering system and Collaborative filtering system are explained in section 3. Proposed system design is explained in section 4. Implementation of the system is presented in section 5. Finally we conclude this system in section 6.

2. Related Works

The recommendation system is often associated with content-based and collaborative filtering systems. System recommendation based on the opinions of like-minded users by collaborative filter and show preference news based on content-based filter. The main advantages of the content-based filtering which include early predictions that cover all items and users, collaborative filtering is capable of accurately recommendation.

CF approaches are often classified as memory-based or model-based. Memory-based use user rating data to compute similarity between users or items. This is used for making recommendations. Based on the rating of similar users or items, recommendation for test user can be generated. The advantages of memory-based approach are the explainability of the result which is an important aspect recommendation system. New data can be added easily and incrementally. In model-based approach, training dates are used to generate a model and to make predictions for real data.

3. Web Personalization System

Web personalization is defined as any action that adapts the information or services provided by a Web Site to the need of a particular user or a set of users, taking advantage of the knowledge gained from the user's navigational behavior and individual interests, in combination with the content and structure of the Web site. The objective of Web personalization system is to "provide user with the information they want or need, without expecting from them to ask for if explicitly".

3.1. Content-Based Filtering

Our content-based filtering algorithms match article keywords to keywords in the user profile. We first briefly describe the format of the user's profile required for the calculation of the content-based prediction, then give detail on keyword generation and lastly describe our matching function.

Each user profile is divided into sections, such as "Business" or "Sport". User can explicitly indicate preference for articles in these section by making the checkboxes for each particular section. In addition, user can specify explicit keywords for each section. Explicit keywords enable the content-based filter to use direct learning in predicting article interest. Direct learning provides predictable behavior for the user and is often precise in defining user interest.

To generate keywords, for each article we remove stop words and then perform word steaming. Keywords are selected based on a frequency count of words. To compute the degree of match between the article keywords and the keywords in the user's profile, we use the Overlap Coefficient given in the following formula:

$$M = \frac{2|D \cap Q|}{\min(|D|,|Q|)}$$

Where D is a set of keywords extracted from the article and Q is the set of keywords in the user's profile. The coefficient, M is not influenced by the size of D and Q, which is desirable as the number of article keyword could be much larger than or smaller then the keywords in the user's explicit keywords list.

3.2. Collaborative Filtering

Collaborative filtering recommendation is the most successful recommendation technique to date. The goal of a collaborative filtering algorithm is to suggest new items or to predict the utility of a certain item for a particular user based on the user's previous linked and the opinions of other liked-minded users. This approach computes recommendations by identifying the user with similar preference and then their items/content ratings are used to propose item/content to one another.

In our system, we used memory-based collaborative recommendation. This algorithm is the most popular prediction technique in CF application. Memory-based algorithms have the advantages of being able to rapidly incorporate the most up-to-date information and relative accurate prediction. The Pearson Correlation Coefficient is widely and successfully used as a similarity measure between users. The new user has an average rated for news he/she has rated. Then the predicted rating of the new user over other products could be calculated by adding weighted sum of other user rating. The weight could be determined by the similarity between active user and other users [2, 6].

According to Pearson Correlation Coefficient step1 can be provide that computing similarity between users as shown in equation1. After, step2 can also proved that calculating the predicted rating of active user for item i as shown in equation2

Step1: Using the Pearson Correlation Coefficient compute similarities between users a and b:

Equation (1)

$$\mathrm{corr}_{a,b} = \frac{\sum_i (r_{ai} - \bar{r}_a)(r_{bi} - \bar{r}_b)}{\sqrt{\sum_i (r_{ai} - \bar{r}_a)^2 \sum (r_{bi} - \bar{r}_b)^2}}$$

Step2: The prediction for item i for user a:

Equation (2)

$$\mathrm{prediction} = \bar{r}_a + \frac{\sum_{i=1}^{n} (\mathrm{corr}_{a,b}) \times (r_{bi} - \bar{r}_b)}{\sum_{i=1}^{n} (\mathrm{corr}_{a,b})}$$

Where, $\bar{r}_a$ is average of current user a's rating for all item, n is the total number of user in the system. $\bar{r}_b$ is the average of other user b's rating for all items. r_{bi} is rating other user a give item. $\mathrm{corr}_{a,b}$ is similarity between user a and b. r_{ai} is rating current user and b.

4. Proposed System Design

In our news personalization system, have four main components: (1) user must login. (2) finding recommended news based on profile by content-based filtering. (3) calculate recommended news by collaborative filtering. (4) display predicted news by combining the content-based filtering with the collaborative filtering algorithm as shown in Figure 1.

When user login, user must fill detail in detail information and preference. After, match keywords in the article with keywords in the user profile by content-based filtering algorithms. To compute the degree of match between the keywords in the article and keywords in the user profile, we use Overlap Coefficient algorithm. Predictions for user from two interest indicator, explicit keywords, and newspaper section, are combined via the matching function with the article keywords and section. If a user has checked only “Sport” in the section, sport articles will be preference. When compute prediction, let two words in the user profile and ten words in the article then inter set these keywords multiply by 2. And find minimum keywords in the articles and keywords in the user profile then format 4/2. Answer is 2. It is degree of Overlap Coefficient.

And then, user can present rating to user interest’s news. Accept user rate from scale and rating scale have 1 to 4 scales. For rating value and rate meaning are

1. I don’t like.
2. It’s ok.
3. I like it.
4. I love it.

This rating is news recommender system to use CF algorithms of memory-based approach. In our system, user can search by categories (type of news) and title (article).

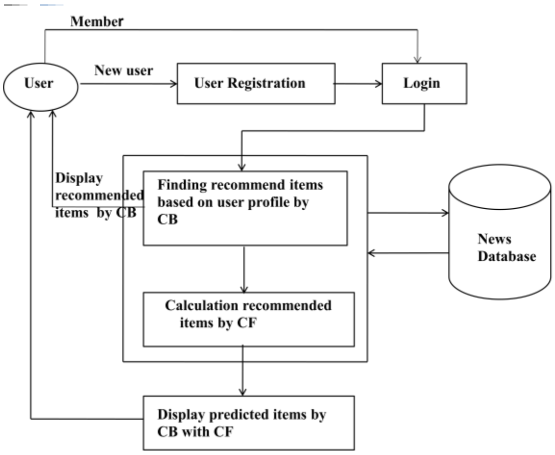

Member
User
New user
User Registration
Login
Display recommended items by CB
Finding recommend items based on user profile by CB
News Database
Calculation recommended items by CF
Display predicted items by CB with CF

Figure 1: Overview of the System Design

In this system needs user profile to give recommendation. The user creates profile by filtering the form and save the user information which provide an easier way to gather accurate information about the users. This system allows new user can free register because user can get recommendation and that information will be used in the process of computing recommendations. If member user can read news and give rate for news, the system can accept rate from user.

And, the system computes weight (similarity) between active user and other users using Pearson Correlation Coefficient algorithm. This algorithm is widely and successful similarity measure between users based on average rating. There are two steps in Pearson Correlation Coefficient. First, get current user and user’s ranked items information. And get similar users and their items. Then calculate the similarity in the equation (1). The result of equation (1) describe in Table 1. Second, calculate prediction for remaining items to give recommend the active user in equation (2).

Step 1: User can rate for items or products like-minded. After got rate value, the system stores the dataset with item. We support there are five users describe in table. User5 is active user. User1, User2, User3 and User4 are other users. There are calculating similarities as follows:

Table 1. Average rate value for news

Item/User	User 1	User 2	User 3	User 4	User 5
news 1	3	1	2	2	3
news 2	2	3	2	1	2
news 3	3	1	1	3	3
AVG	2.67	1.67	1.67	2.00	2.67

(User 1, User 5) = 1
(User 2, User 5) = -1
(User 3, User 5) = -0.5
(User 4, User 5) = 0.8

Step 2: After computing similarity between active user and other users, the equation (2) is used to predict recommendation of active user (user5). Correlation coefficient of -1 means no correlation. New user needs rank an item once at time.

Table 2. Average rate for each item

Item/ user	User 1	User 2	User 3	User 4	User 5
news 1	3	1	2	2	3
news 2	2	3	2	1	2
news 3	3	1	1	3	3
news 4	1	2	3	4	?
news 5	4	3	2	1	?
AVG	2.6	2.00	2.00	2.2	2.67

P(User 5, news 4) = 2.2264
P(User 5, news 5) = 2.8867

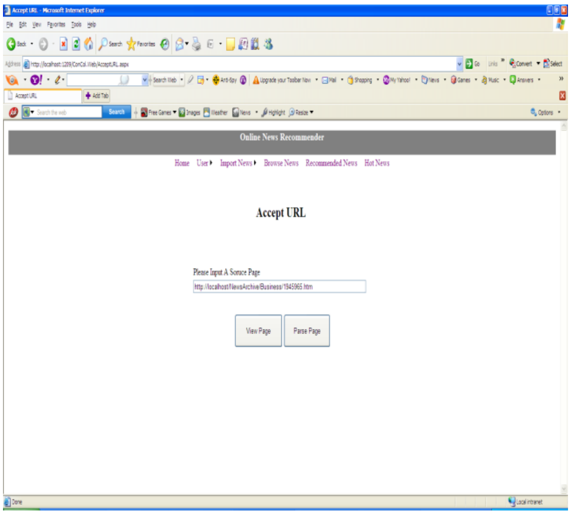
According to Table 2, predict rate all items are computed using equation (2) and shows recommended news for active user. In our system, shows recommended news for active user sorting as ranks by combine degree of overlap coefficient and system ranks. In this system, user can search by news categories according to Table 3.

Table 3. Types of categories

No	Categories
1	Advertising
2	Business
3	Business Opportunity
4	Education
5	Entertainment
6	Fashion
7	Food
8	Health
9	Music
10	Science &Technology
11	Sport

5. System Implementation

In our proposed system, we use news from www.amazines.com newspaper web site and use 10 users and newspaper about fives day. When import the news web page, input the URL link shown in Figure 2 and parse the URL link (.html) of the web page where parse the title, keyword, description and then store the HTML file in the database. We can see HTML file of the web page in Figure 3 which contain title, keywords and description. In this system, to generate keywords, we remove stop words for each article and then perform steaming. When we remove stop words, define the stop words (a, an, the, and, for, from, on, etc) and then put the string array. And remove the stop words in the query as define the stop words. We remove stop words in the keywords and title.



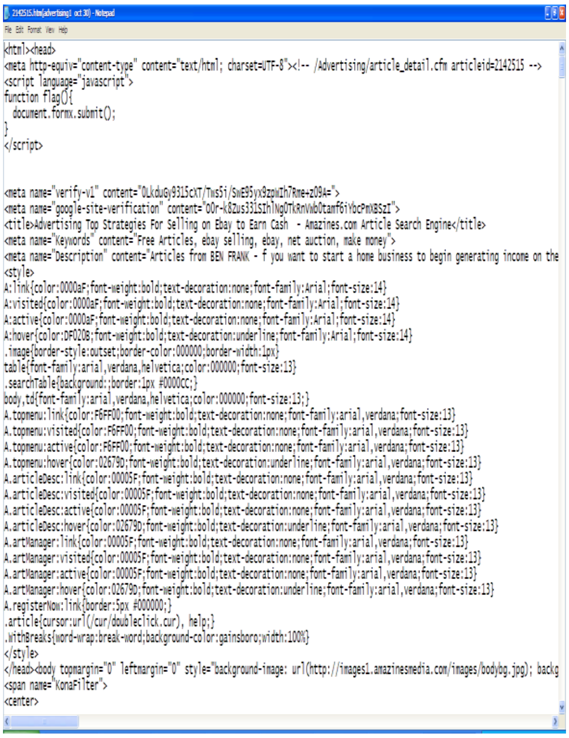
Online News Recommender

Accept URL

Please Input A Soruce Page
http://localhost/NewsArchive/Business/1945965.htm

Accept URL - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://localhost:1209/ConCol.Web/AcceptURL.aspx
Home
User
Import News
Browse News
Recommended News
Hot News
View Page
Parse Page

Figure 2: Input the URL link of Web page



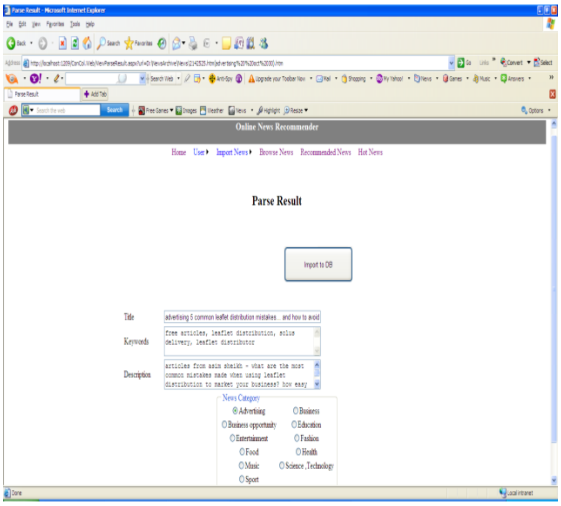
2142515.htm(advertising1 oct 30) - Notepad
File Edit Format View Help

```
<html><head>
<meta http-equiv="content-type" content="text/html; charset=UTF-8"><!-- /Advertising/article_detail.cfm articleid=2142515 -->
<script language="javascript">
function flag(){
  document.formx.submit();
}
</script>



<meta name="verify-v1" content="0LkduGy9315cXT/Tws5i/SwE95yx9zpwIh7Rme+z09A=">
<meta name="google-site-verification" content="O0r-k8Zus331SIhlNgOTkRnVwb0tamf6iYbcPmXBSzI">
<title>Advertising Top Strategies For Selling on Ebay to Earn Cash  - Amazines.com Article Search Engine</title>
<meta name="Keywords" content="Free Articles, ebay selling, ebay, net auction, make money">
<meta name="Description" content="Articles from BEN FRANK - f you want to start a home business to begin generating income on the
<style>
A:link{color:0000aF;font-weight:bold;text-decoration:none;font-family:Arial;font-size:14}
A:visited{color:0000aF;font-weight:bold;text-decoration:none;font-family:Arial;font-size:14}
A:active{color:0000aF;font-weight:bold;text-decoration:none;font-family:Arial;font-size:14}
A:hover{color:DF020B;font-weight:bold;text-decoration:underline;font-family:Arial;font-size:14}
.image{border-style:outset;border-color:000000;border-width:1px}
table{font-family:arial,verdana,helvetica;color:000000;font-size:13}
.searchTable{background:;border:1px #0000CC;}
body,td{font-family:arial,verdana,helvetica;color:000000;font-size:13;}
A.topmenu:link{color:F6FF00;font-weight:bold;text-decoration:none;font-family:arial,verdana;font-size:13}
A.topmenu:visited{color:F6FF00;font-weight:bold;text-decoration:none;font-family:arial,verdana;font-size:13}
A.topmenu:active{color:F6FF00;font-weight:bold;text-decoration:none;font-family:arial,verdana;font-size:13}
A.topmenu:hover{color:02679D;font-weight:bold;text-decoration:underline;font-family:arial,verdana;font-size:13}
A.articleDesc:link{color:00005F;font-weight:bold;text-decoration:none;font-family:arial,verdana;font-size:13}
A.articleDesc:visited{color:00005F;font-weight:bold;text-decoration:none;font-family:arial,verdana;font-size:13}
A.articleDesc:active{color:00005F;font-weight:bold;text-decoration:none;font-family:arial,verdana;font-size:13}
A.articleDesc:hover{color:02679D;font-weight:bold;text-decoration:underline;font-family:arial,verdana;font-size:13}
A.artManager:link{color:00005F;font-weight:bold;text-decoration:none;font-family:arial,verdana;font-size:13}
A.artManager:visited{color:00005F;font-weight:bold;text-decoration:none;font-family:arial,verdana;font-size:13}
A.artManager:active{color:00005F;font-weight:bold;text-decoration:none;font-family:arial,verdana;font-size:13}
A.artManager:hover{color:02679D;font-weight:bold;text-decoration:underline;font-family:arial,verdana;font-size:13}
A.registerNow:link{border:5px #000000;}
.article{cursor:url(/cur/doubleclick.cur), help;}
.WithBreaks{word-wrap:break-word;background-color:gainsboro;width:100%}
</style>
</head><body topmargin="0" leftmargin="0" style="background-image: url(http://images1.amazinesmedia.com/images/bodybg.jpg); backg
<span name="KonaFilter">
<center>
```

Figure 3: HTML page of the news web page

To import the news web page, administrator distinguishes the categories such as Business, sport, advertising, etc. First, accept the URL link (HTML file) of news web page by clicking the "Parse page" button. After accepting the URL link, parse the title, keyword and description of the news web page. And administrator chooses the news categories (such as Business, Sport, advertising, etc and then store news web page in the database shown in the Figure 4. Administrator performs to import the news in the database. Administrator must know kind of news categories (such as Business, sport, etc) to import.

When show the news to the user by categories such as Business, Sport, Advertising, etc. So user can search easily and read news by categories. This system allows users to login, can modify their user profiles, and browse, read and rate newspaper articles. If user is member, the user will come to system login by login Id and password.


Parse Result - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address
Go Links Convert Select
Search Web Anti-Spy Upgrade your Toolbar Now Mail Shopping My Yahoo! News Games Music Answers
Parse Result
Add Tab
Search the web Search Free Games Images Weather News Highlight Resize Options
Done
Local intranet

http://localhost:1209/ConCol.Web/NewParseResult.aspx?url=D:\NewsArchive\News\2142525.htm\advertising%20%20oct%2030\.htm

Online News Recommender

Home User Import News Browse News Recommended News Hot News

Parse Result

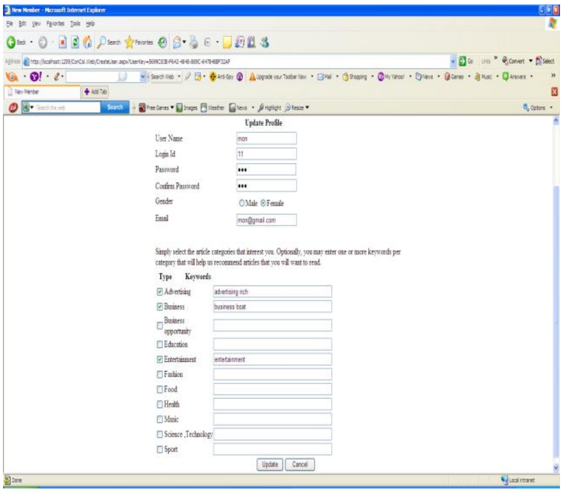
Import to DB

Title	advertising 5 common leaflet distribution mistakes... and how to avoid
Keywords	free articles, leaflet distribution, solus delivery, leaflet distributor
Description	articles from asim sheikh - what are the most common mistakes made when using leaflet distribution to market your business? how easy

News Category

● Advertising ○ Business
○ Business opportunity ○ Education
○ Entertainment ○ Fashion
○ Food ○ Health
○ Music ○ Science ,Technology
○ Sport

Figure 4: Import the News to database


New Member - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address
Go Links Convert Select
Search Web Anti-Spy Upgrade your Toolbar Now Mail Shopping My Yahoo! News Games Music Answers
New Member
Add Tab
Search the web Search Free Games Images Weather News Highlight Resize Options
Done
Local intranet

Update Profile

User Name	mon
Login Id	11
Password	•••
Confirm Password	•••
Gender	○ Male ● Female
Email	mon@gmail.com

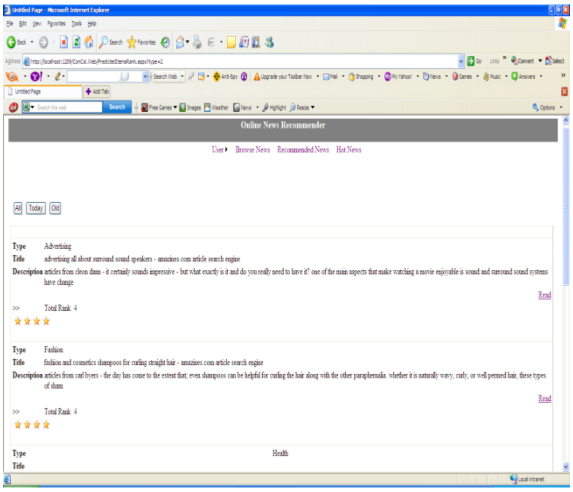
Simply select the article categories that interest you. Optionally, you may enter one or more keywords per category that will help us recommend articles that you will want to read.

Type	Keywords
☑ Advertising	advertising rich
☑ Business	business boat
□ Business opportunity	
□ Education	
☑ Entertainment	entertainment
□ Fashion	
□ Food	
□ Health	
□ Music	
□ Science ,Technology	
□ Sport	

Update Cancel

Figure 5: User Profile. User can change information in This Page

User can see preference news by content-based filtering gives the rate, see recommended news and hot news. If user is not member, allow users see browse news and go to registration, fill detail information and preference. After registration, user can be seen recommended news and hot news by content-based with collaborative filtering. Member user can be modifying their profile.


Untitled Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address
Go Links Convert Select
Search Web Anti-Spy Upgrade your Toolbar Now Mail Shopping My Yahoo! News Games Music Answers
Untitled Page
Add Tab
Search the web Search Free Games Images Weather News Highlight Resize Options
Local intranet

http://localhost:1209/ConCol.Web/PredictedItemsRank.aspx?type=2

Online News Recommender

User Browse News Recommended News Hot News

All Today Old

Type Advertising
Title advertising all about surround sound speakers - amazines.com article search engine
Description articles from cleon dann - it certainly sounds impressive - but what exactly is it and do you really need to have it? one of the main aspects that make watching a movie enjoyable is sound and surround sound systems have change

Read

>> Total Rank 4
★★★★

Type Fashion
Title fashion and cosmetics shampoos for curling straight hair - amazines.com article search engine
Description articles from carl byers - the day has come to the extent that; even shampoos can be helpful for curling the hair along with the other paraphernalia. whether it is naturally wavy, curly, or well permed hair, these types of sham

Read

>> Total Rank 4
★★★★

Type Health
Title

Figure 6: Recommended News Page

The user profile page lists the available newspaper section and provides checkboxes for indication of interest in each section as shown in Figure 5. In addition user can specify multi-word explicit keyword within each section. The newspaper sections supported are: Advertising, Business, Business opportunity, Education, Entertainment, Fashion, Food, Health, Music, Science and Technology, and Sport. Recommended news and hot news are displayed for user by calculating prediction. User can change their information. And then news articles are displayed in Figure 6 by sorting as rank with content-based and collaborative filtering. When show recommended news, user can read today news, old news and all. Show recommended news with two stars and above and then show hot news with three stars and above. Value of ranks represent stars symbol.

6. Conclusion

Recommender system has emerged as powerful tools for helping users find and evaluate items of interest. The presented scheme uses a content-based predictor to enhance accurate preference of user and then provide personalized suggestions through collaborative filtering. The news is recommended by using content-based and collaborative filtering. Our system use memory-based approach that reduce and get extensive information. This approach used in internet application such as laptop, music, book store, web page, etc. This system information provides the users with most effective and efficient information that is relevant to the user’s request.

7. References

[1] Cyrus Shahabi and Yi-Shin Chen, “Web Information Personalization : Challenges and Approaches”, Department of Computer Science, University of Southern California, Los Angeles, CA 90089-2561, USA.

[2] G. Linden, B. Smith and J. York, “Amazone.com Recommendation: Item-to-Item Collaborative Filtering”, Industry Report from IEEE Internet Computing.

[3] Mark Claypool, Anuja Gokhale, Tim Miranda, Pavel Murnikov, Dmitry Netes and Matthew Sartin, “Combining Content-Based and Collaborative Filters in an Online Newspaper”, Computer Science Department, Worcester Polytechnic Institute Worcester, Msssachusetts, USA.

[4] Nave Tintraev and Judith Masthoff “Similarity for News Recommender Systems”, University of Aberdeen, Aberdeen, UK.

[5] Sarabjot Singh Anand and Bamshad Mobasher, “Intelligent Techniques for Web Personalization”, Department of Computer Science, University of Warwick, Coventry CV47AL, UK.

[6] Y. Qu, X. Yang, T. Haung “Survay of Recommender Systems and Algorithms”, EE380L: Data Mining.