Enhancing Collaborative Filtering Based Recommender System by Using Adaptive Clustering

J. Bhavithra[1], Dr. A. Saradha[2], K. Jayanthi[3]

[1]Assistant Professor, [3]Student, Dept. of CSE,
Dr. Mahalingam College of Engg. & Tech., Pollachi, TamilNadu, India
[2]Associate Professor and Head, Dept. of CSE, IRTT, Erode, Tamilnadu, India

Abstract -A recommender system is an information system that suggests items, web pages to a web user. Collaborative filtering based recommender system recommends the list of web pages to the user based on other similar user's preferences. Newly created web pages arriving to the search engine will not be considered for recommendation, as it is not visited by any users. This is termed as cold start problem. This work includes semantic similarity based relationship as an add-on to the frequency of keywords in the visited web pages. The work also uses adaptive clustering mechanism in order to cluster all the frequent keywords based on their relationship. When the user enters a query, the web pages that contain the keywords that are semantically similar to the top keywords in the matching cluster corresponding to query will be recommended to the user. The current recommender system considers not only the popular web pages, but also every newly created and dynamically modified web pages. Thus avoiding cold start problem and popularity bias. The system has been tested by providing single-keyword queries and the results are compared with existing collaborative based recommender systems. It has been observed that, the accuracy of the Semantic similarity based adaptive clustering recommendation technique has been increased by 20% comparing to existing system. Also the recommendation diversity has been increased by 11% to that of existing system.

Keywords - Cold start problem, semantic similarity, clustering mechanism, Normal Recovery Collaborative Filtering, Keyword Density.

I. Introduction

One of the applications of data mining technique is web mining, it determine patterns from online. The Recommender system is a one of the part of web mining. Recommender systems are typically producing a list of web page recommendation to user. This system is an information filtering system, which predicts the preference from the past users [1]. Recommender systems typically produce a list of recommendations in one of two ways through collaborative and content-based filtering. Collaborative filtering approaches make a model from a user's past behaviour (previously visited web pages) as well as similar decisions made by other users. The users repeatedly visited web pages are recommended to next users by using the user's preferences. The users browsing history has been collected and it is called as web log file. The log file has users id, user visited web page for user's particular query and time of visiting the web pages. User's profiles will be generated from the web log file. User's profile has a user's web page navigation behaviour and keywords are represented.

II. Normal Recovery Collaborative Filtering

Similarity measures are performed between users by using the approach Normal Recovery Similarity Measure. Normal Recovery method has been proposed by Huifeng Sun Et. Al (2013). Here, the similarities are measures for nearest neighbours (K-NN) [Huifeng Sun Et. Al (2013)]. The two users have similar profiles

 ISBN : 978-81-929742-0-0

then the users visited web pages are recommended to another new user. Similarly the lists of web pages are recommended to users.

$$sim(u,v) = 1 - \frac{\sqrt{\sum_{i \in I} \left(\frac{r_{u,i} - r_{umin}}{r_{umax} - r_{umin}} - \frac{r_{v,i} - r_{vmin}}{r_{vmax} - r_{vmin}}\right)^2}}{\sqrt{|I|}} \quad (1)$$

Similarities between the two users were calculated by using the formula (1) [Huifeng Sun Et. Al (2013) [3]]. Where i is the set of web pages that are co-visited by user u and v. |I| is the number of I, i.e. total number web pages co-visited by users u and v. $r_{u,i}$ is the value of web page keyword and time spent in particular web page from user u in user web page matrix. r_{umin} and r_{umax} are the lowest and highest values of user u. r_{vmin} and r_{vmax} denotes the lowest and highest values of user v. The similarity is measured from the equation. The sim(u,v) are in values between 0 to 1. Eq (1) has been used for calculating the similarity between two users. The NRCF approach [Huifeng Sun Et. Al (2013) [3]] can adapt to different environment easily.

$$\hat{r}_{u,i} = \lambda \times r_{umin} + (r_{umax} - r_{umin}) \frac{\sum_{u' \in U} Sim(u,u') \times nr_{u',i}}{\sum_{u' \in U} Sim(u,u')} \quad (2)$$

Eq(2). [Huifeng Sun Et. Al (2013) [3]] is used for finding the web service recommendation to users. Where λ = 1, then the recommender system make recommendation to user, because the users are similar. r_{umin} and r_{umax} are the lowest and highest values of user u. $Sim(u,u')$ are the values of similarity between two different users. $nr_{u',i}$ Is an average value of user u (keywords and time spend in a web page). After calculating the value of recommendation approach compare the another users values and recommend the most visited web pages to user.

The collaborative filtering approach recommends only the most frequently visited web pages by web users. The newly created web pages or modified web pages will not be considered effectively for recommendation to the current user. This is stated as cold start problem. Even dynamic web pages, where the keyword changes frequently, are also not considered effectively for recommendation to active users. To avoid this problem the semantic similarity between web pages and clustering mechanism is introduced along with collaborative filtering mechanism.

This paper proceeds as follows. Section III describes the semantic similarity between the keywords. The cluster formation (adaptive cluster) is presented in Section IV. Section V presents semantic similarity based recommendation process. Section VI explains the result and finally VII explains about calculation and future work.

A. Dataset

The data are collected from the user browsing history through AOL search data. AOL dataset is anonymized form of web server search log. This dataset is downloaded from [17]. Around 650K users and 8, 00,000 URLs were collected. The user browsing date and time, user's searched query and user visited URLs are collected. The collected dataset is in the form of text file. The user's identity was anonymized to a random id in the downloaded data set. Each user has an id number; the scale of the dataset is very large in the field of web service recommendation. The keywords are extracted from the user visited web page contents. The web pages are also having many advertisements and etc. The advertisements are removed and get only get the keyword of the web pages.

III. Semantic Similarity in Dynamic Web Pages

The semantic similarity is more suitable for finding the relationship between the keywords existing in the web page. It finds how much, the two keywords are matched, and hence the relationship between any two web pages. The web resources, including webpages and contents, and hence keywords, are dynamic in

nature due to the development and periodic changes on the contents. This dynamism in the content is indeed mandatory to be considered for web page recommendation process. The keywords are considered as an important aspect for finding the semantic similarity between these dynamic web pages. After extracting the all keywords from the user visited web pages, the semantic similarity between those keywords are obtained. The similarities between the keywords will predict the relationship between the active user session and user's intended web pages which are to be recommended.

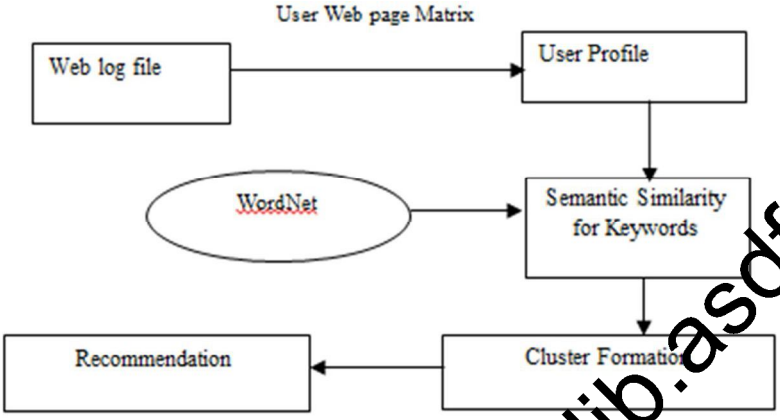

Fig 1: Semantic similarity based recommendation

RiTa WordNet is a library that provides simple access to the WordNet ontology for language-oriented operations. This paper uses RiTa WordNet for finding the semantic relationship between the keywords. The value 0.0 represents the keywords that are more similar and 1.0 represents the keywords that are dissimilar.

The similarities of those keywords identified in user accessed web pages (obtained through web logs) are computed. A matrix as shown in Table I is populated by comparing all possible keywords in all web pages.

Table1: Semantic similarity between keywords

Keywords	Apple	Electrical	Machine	Mobile	Text
Apple	0.0	0.97	0.76	0.11	1.0
Electrical	0.97	0.0	0.23	0.44	1.0
Machine	0.76	0.23	0.0	0.18	1.0
Mobile	0.11	0.44	0.18	0.0	0.09
Text	0.25	1.0	1.0	0.09	0.0

IV. Cluster Formation

Clustering, a process of grouping keywords, is performed on those set of keywords in table I. Based on relationship values between the keywords, the cluster will be formed. For each keyword listed in table 1, two possible clusters are resulted. Group A contains keywords that are above the median value and Group B with keywords that are below the median value. As an example, consider the first row values of table I. Median is computed for all related keyword measures in row 1. For keyword "Apple" (row 1) the minimum value is 0.0 and maximum is 1.0. Take a median for minimum and maximum values (0.0 + 1.0 / 2 = 0.5). The keyword whose similarity is below median (0.0 to 0.5) is clustered into Group B. Those keywords whose similarity is greater than median (0.51 to 1.0) will be clustered as Group A. Words in Group B are considered as more similar and used for further recommendations. The group B of keywords will be maintained inside user's machine as a cookie file. The keywords and group will be updated periodically during further searches.

V. Semantic Similarity Based Recommendation

In recommendation process, new web pages are recommended to user based on user's query and cluster (group) match. Initially, during an active session, user is allowed to enter the query. The query is then passed to search engine as usual. From the initial search engine's result, all the web pages resulted are mined for selecting the "apt" keywords. "Apt" keywords are termed as those which are in par with the context of the search, after removing stop words and duplicates. For each "apt" keyword further check is done for matching it with Group B of the corresponding user (stored as cookie file). Keyword density can also be used as a factor in determining "apt" keywords and matching ones in clusters. If the keywords are available in cluster (group B), then the corresponding cluster's containing keyword's web pages are recommended to user. Hence, recommending the web pages that are more accurate, more diversified in popularity and likely to be visited by the active user.

VI. Result

The recommendation diversity is found to be high comparing to existing systems. As the current system considers both unvisited web pages and freshly created/ updated web pages, diversity is enhanced and cold start problem is avoided.

Table 2: Comparison between systems

Keywords / Single word Query	Number of relevant web sites recommendation		Diversity of web pages recommendation
	Existing System	Proposed System	
Company	3	11	11
Web	5	10	8
Network	2	12	12
Birthday	5	9	9
Email	4	11	10
News	1	14	14

The above table 2 compares the recommendation of relevant web pages in existing and current system. Single word query recommendation is high, because the current system avoids the cold start problem and considers dynamism of web pages. Comparing to existing system, the current system recommends more number of web pages to web user. Both users visited and unvisited web pages are recommended in current system.

Recall, precision and F-measure metrics were calculated using single word query recommendation and compared with existing system. The accuracy of semantic similarity based collaborative recommender system has been improved by 20% comparing to existing system.

$$\text{Precision} = \frac{|Ra|}{|A|} \quad (3)$$

$$\text{Recall} = \frac{|Ra|}{|R|} \quad (4)$$

$$\text{F-Measure} = \frac{2}{\frac{1}{r(j)} + \frac{1}{p(j)}} \quad (5)$$

Table 3: Accuracy Comparison

Keywords	Recall		Precision		F-Measure	
	Existing	Current	Existing	Current	Existing	Current

Company	20	73	25	91	66	81
Web	13	66	16	83	14	41
Network	26	53	33	66	29	58
Birthday	20	60	25	75	66	66
Email	33	80	41	100	36	88
News	26	80	33	100	29	90
wireless	6	55	12	66	38	60
Market	14	73	18	91	14	81
Share	35	82	40	88	37	89
Business	29	60	38	75	30	63

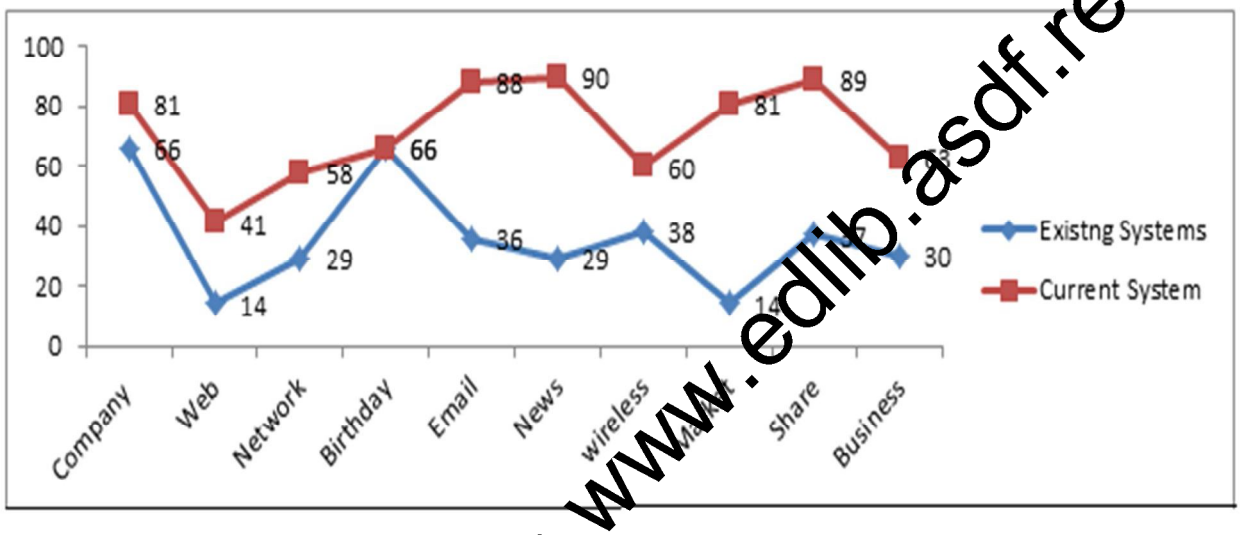

Fig 2: Accuracy graph

Fig 2 shows that accuracy graph compares the accuracy level of existing and current system. Based on single word query recommendation, the newly created, dynamic web pages are also considered for recommendation. Diversity of recommendation has been improved by 11%. Hence accuracy has been enhanced by 20%.

VII. Conclusion and Future Work

In this paper, a recommendation method that is based on semantic similarity between keywords has been proposed which is used to recommend the new and dynamic web pages to user. The proposed approach also uses adaptive clustering mechanism in order to cluster all the frequent keywords based on their relationship. The proposed recommender system considers not only the popular web pages, but also every newly created and dynamically modified web pages. Thus avoiding cold start problem and popularity bias. Proposed system has been tested by providing single-keyword queries and the results are compared with existing collaborative based recommender systems. It has been observed that, the accuracy of the proposed Semantic similarity based adaptive clustering recommendation technique has been increased by 20% comparing to existing system. Also the recommendation diversity has been increased by 11% to that of existing system.

As a future work of this paper, the recommendation can still be optimized based on context based approaches. Instead of single keyword queries, the current system can be extended to include phrased queries.

References

1. Mcilraith, Sheila A., Tran Cao Son, and Honglei Zeng, (2012), "Semantic Web Services", IEEE Transactions on Intelligent Systems, Volume No: 2, pp. 46-53.
2. Zheng, Zibin, Hao Ma, Michael R. Lyu, and Irwin King, (2009), "Wsrec: A Collaborative Filtering based web service Recommender System", International Conference on IEEE, pp. 437 - 444
3. Huifeng Sun, Zibin Zheng, Junliang Chen, and M. R.Lyu, (2013), "Personalized Web Service Recommendation via Normal Recovery Collaborative Filtering", IEEE Transactions on Services Computing, Volume 6 No:46, pp.573 - 579
4. Zheng, Zibin, Hao Ma, Michael R. Lyu, and Irwin Kin, (2011), "QoS-aware Web Service Recommendation by Collaborative Filtering", IEEE Transactions on Services Computing, Volume No: 4, pp.140-152.
5. Hwang, Cheinshung, and Tungsheng Chang, (2012), "Genetic K-Means Collaborative Filtering for Multi-Criteria Recommendation", Journal of Computational Information Systems, Volume No: 8, pp.293-303.
6. Deshpande, Mukund, and George Karypis, (2004), "Item-Based Top-N Recommendation Algorithms", ACM Transactions on Information Systems (TOIS), Volume No: 22, pp.143-177.
7. Sun, Huifeng, Yong Peng, Junliang Chen, Chuanchang Liu, and Yuzhuo Sun, (2011), "A New Similarity Measure Based on Adjusted Euclidean Distance for Memory-Based Collaborative Filtering", Journal of Software, Volume No: 6, pp.993-1000.
8. Takale, Sheetal A., and S. Nandgaonkar, (2010), "Measuring Semantic Similarity between Words using Web Documents", International Journal of Advanced Computer Science and Applications-IJACSA, Volume No:1, pp.78-85.
9. Howe, Daniel C. (2009), "RiTa: Creativity Support for Computational Literature", In Proceedings of the Seventh ACM conference on Creativity and Cognition, pp.205-210.
10. Abhishek, Vibhanshu, and Kartik Hosanagar, (2007), "Keyword Generation for Search Engine Advertising using Semantic Similarity between Terms", Proceedings of the Ninth International Conference on Electronic Commerce on ACM, pp.89-94.
11. Symeonidis, Panagiotis, Alexandros Nanopoulos, Apostolos Papadopoulos, and Yannis Manolopoulos, (2007), "Nearest-Biclusters Collaborative Filtering with Constant Values", Advances in web mining and web usage analysis on Springer, pp. 36-55.
12. Park, Yoon-Joo, (2013), "The Adaptive Clustering Method for the Long Tail Problem of Recommender Systems", IEEE Transactions on Knowledge and Data Engineering, Volume No: 25, pp.1904-1915.
13. Gong, Songjie, (2010), "A collaborative filtering recommendation algorithm based on User Clustering and Item Clustering", Journal of Software, Volume No: 5, pp.745-752.
14. Pakhira, Malay K, (2009), "A modified K-Means Algorithm to avoid Empty Clusters", International Journal of Recent Trends in Engineering, Volume No:1, pp.220-226.
15. Gunduz-Oguducu, Şule, (2010), "Web Page Recommendation Models: Theory and Algorithms", Synthesis Lectures on Data Management Volume No: 2, pp.1-85.
16. AOL Dataset, "http://www.michael-noll.com/blog/2006/08/07/aol-research-publishes-500k-user-queries/"
17. Dataset, AOL Search Data, http://www.infochimps.com/datasets/aol-search-data