Personalized News Recommendation Based on Consumers' Click Behavior

Yuqi Wang
School of Computer Science
Communication University of China
Beijing, China

Wenqian Shang
School of Computer Science
Communication University of China
Beijing, China

Abstract—**The news browsing sequence of a consumer can be obtained from the consumer's click behavior on the Internet. Here, some potential associations between news using the news browsing sequence of a consumer will be found. Then, personalized news recommendation for different consumers can be provided according to these potential associations. In this paper, an improved personalized news recommendation algorithm based on consumers' click behavior is proposed. Through doing experiments on real news browsing data, the recommendation result is better and the new algorithm is proved to be feasible.**

Keywords-personalized news recommendation; association rules; click behavior

I. INTRODUCTION

With the development of the Internet, more and more people read news on the Internet. However, facing the massive data, consumers need to spend much time on choosing news they prefer. Therefore, providing an effective personalized news recommendation for consumers will promote consumers' experience of browsing news greatly [1].

The research on the personalized recommendation system can be roughly divided into four kinds, namely recommendation based on content, recommendation based on collaborative filtering, recommendation based on knowledge and hybrid recommendation [2]. The recommendation based on content is a method which is widely used in practice. This method finds the relation between different news according to the content. Then, the preference of consumers can be obtained and applied to news recommendation. As text processing technology gets more sophisticated, the research on the recommendation based on content is becoming more and more mature. The recommendation based on collaborative filtering is a method which is widely studied in recommendation system. This method has two branches, namely collaborative filtering based on items and collaborative filtering based on consumers [3]. The collaborative filtering based on items obtains similarity of different items from their scores provided by same consumers and make recommendations according to the similarity of items. The collaborative filtering based on consumers obtains similarity of different consumers from their evaluation of the same items and make recommendations according to the similarity of consumers. The recommendation based on collaborative filtering can be implemented easily and provide accurate recommendations regardless of the forms of content [4]. But, this method is hard to provide a good recommendation for new items and new consumers because of the lack of information of evaluation. This problem is called cold-start [5]. The recommendation based on knowledge will be used when there is not enough recommendation information provided. So, this method can solve cold-start problem well. The knowledge includes consumers' preference, product knowledge, functional knowledge and so on [6]. This information is independent of historical behavior data of consumers. But, in the process of recommending news, such knowledge is difficult to obtain because of uncertainty of consumers. The hybrid recommendation method is a combination of the above three methods. In most cases, each recommendation algorithm has some shortcomings and limitations [7]. So, in order to obtain more accurate recommendation results, the integration of different methods is necessary. Actually, most recommendation algorithm applied in recommendation systems is hybrid recommendation.

There are many research about the recommendation algorithm have done based on the above four methods. M. Nakagawa proposes a method to mine consumers' click behavior from website logs by using association rules [8]. Consumers browse the website according to their interests, so there will be some potential association between the pages clicked by same consumers. And the recommendation is made according to strong rules mined by frequency items [9, 10]. However, association rule encounters the limitations including scalability and efficiency because of a large amount of calculations [11]. O. Nasraoui proposes a method to make recommendation by obtaining the consumer's information combined with the fuzzy of ANN [12]. They group the profiles using unsupervised hierarchical clustering. M. Awad proposes a method to improve prediction accuracy by combining domain knowledge with VSM and MM [13, 14]. However, it costs too much training time because of a large number of tags and categories. M. T. Hassan proposes a method to predict the consumer's current profile by Bayesian network [15, 16]. Bayesian network takes full advantage of the priori information and the time sequence for recommendation. Speretta proposes a method to create consumers profiles through classifying queries and snippets of clicked search results [17]. Tan presents a recommendation system based on semantic, in which the consumer's profile is represented as weighted concept vectors and several similarities are measured by calculating the

978-1-4673-7682-2/15/$31.00 ©2015 IEEE

similarities between articles [18]. The above algorithms can obtain good recommendation result, but there are limitations on the operation efficiency.

In this paper, a new recommendation algorithm based on collaborative filtering is proposed. Because consumers' preferences are the basis for collaborative filtering [19], it should be obtained. But the news is different from commodity, consumers' evaluation on news cannot be obtained directly. Here, the preferences are obtained through potential associations between news in the news browsing sequence of a consumer.

The rest of this paper is organized as follows. Firstly, the analysis of consumers' click behavior is introduced in section II, and then the personalized news recommendation algorithm is proposed in section III. In section IV, the experiments and analysis are presented. Finally, a conclusion will be given in section V.

II. THE ANALYSIS OF CONSUMER'S CLICK BEHAVIOR

A piece of news is browsed accompanying with consumers' click behavior. So, getting the information of consumers' browsing behavior through the analysis of the consumers' click behavior directly is possible. This information consists of two parts, namely news id and browsing time. Here, the following table can represent a consumer's click behavior well.

TABLE I. CONSUMER'S CLICK BEHAVIOR

Consumer ID	*News ID*	*Browsing Time*

Firstly, analysis on the characteristics of news is made. The news in the Internet is presented on web pages, so the association between news is affected by the link structure of web pages. And consumers' browsing behavior is usually like this: firstly, a consumer begins browsing behavior with a base web page, usually called homepage, and then the consumer will select a piece of news from homepage to continue browsing behavior. Sometimes, the consumer will return to the previous page to select other pages to browse. Through repeating such behavior, the news browsing sequence of the consumer is formed. This characteristic of the consumer's click behavior provides a good idea to obtain the consumer's preference. Here, the distance on the link structure of web pages can be used to represent the degree of correlation between web pages and account that there is correlation between neighboring news in the news browsing sequence of a consumer. So, this potential correlation to find the consumer's preference can be used.

Then, analysis on the browsing time is done. The consumer's click behavior has timeliness and is not always continuous. Neighboring news in the news browsing sequence of a consumer may be browsed on different days and the correlation between them is not very reliable. The browsing time difference between neighboring news should be considered. Then, there is another question, namely how to set the upper and lower limits of the browsing time difference. Actually, the limits of the browsing time difference are difficult to determine and the browsing time difference is needed to promote the confidence of correlation between neighboring news. So, some experiments are necessary to determine the value of limits.

Consumers' click behavior is analyzed from two aspects. Traditional news recommendation algorithm just considers the characteristics of news but ignores their limitations. So, this method will produce many low association rules. Generally, the precision of recommendation based on traditional news recommendation algorithm is low because of the negative influence of low association rules. In this paper, the limitation of browsing time difference is added into the construction of association rules to get a better result.

III. THE PERSONALIZED NEWS RECOMMENDATION ALGORITHM

A. Traditional Algorithm

Based on the above analysis, building association rules between neighboring news in the news browsing sequence of a consumer firstly is necessary. Then, news recommendation for a consumer is made according to the last browsing news. Here, let N represent the total number of consumers and $U_i(n_1, n_2, \cdots, n_k)$ represent the news browsing sequence of the consumer U_i. So, the association rules between neighboring news can be represented by the following formula:

$$n_i \Rightarrow n_{i+1}. \quad (1)$$

Then, let $\langle n_i, n_{i+1} \rangle$ represent a newsgroup with relevance. And the following formula is used to calculate the support of association rules:

$$\text{support}(n_i \Rightarrow n_{i+1}) = \frac{count(\langle n_i, n_{i+1} \rangle)}{N}. \quad (2)$$

where $count(\langle n_i, n_{i+1} \rangle)$ represents the number of consumers whose news browsing sequence contains the newsgroup. According to the support, the value of confidence can be calculated. The formula is given as follows:

$$confidence(n_i \Rightarrow n_{i+1}) = \frac{\text{support}(n_i \Rightarrow n_{i+1})}{\text{support}(n_i)}. \quad (3)$$

Based on the above steps, we can obtain the final recommendation result. And there are some strategies provided to choose. In this paper, the following method is used as the strategy for obtaining final recommendation result.

There are many newsgroups which will be found from data set. But most of them are useless for recommendation. So data screening is important. Let $MinConf$ represent threshold of

confidence. If $\langle n_i , n_{i+1} \rangle$ is useful for recommendation, the following formula must be established:

$$confidence\left(n_i \Rightarrow n_{i+1}\right) \geq MinConf. \quad (4)$$

After obtaining the association rules of news browsing sequence, the news recommendation for each user can be made according to his or her click behavior.

B. Improvement

The above steps implement the traditional news recommendation algorithm. Then, some improvement for it will be done. There are a lot of information be contained in users' click behavior. So, mining users' click behavior can bring improvement of the effect of algorithm. The limitations of browsing time difference can promote the reliability of association rules. Here, let $time\left(n_i\right)$ represent the browsing time of news n_i and $subTime\left(\langle n_i , n_{i+1} \rangle\right)$ to represent the browsing time difference of $\langle n_i , n_{i+1} \rangle$. The following formula can calculate the browsing time difference:

$$subTime\left(\langle n_i , n_{i+1} \rangle\right) = time\left(n_{i+1}\right) - time\left(n_i\right). \quad (5)$$

Let T_h represent the upper limit and T_l represent the lower limit. The formula (5) represents the ranges of the browsing time difference as follows:

$$T_l \leq subTime\left(\langle n_i , n_{i+1} \rangle\right) \leq T_h. \quad (6)$$

Based on the parameters settings of traditional algorithm, $U_i\left(n_1, n_2, \cdots, n_k\right)$ represents the news browsing sequence of the consumer U_i. There will be a sequence value calculated through the above steps. Here, let $Seq\left(v_{\langle n_1, n_2 \rangle}, v_{\langle n_1, n_3 \rangle}, \cdots, v_{\langle n_{k-1}, n_k \rangle}\right)$ represent the sequence value. And $v_{\langle n_i , n_j \rangle}$ represents time difference of $\langle n_i , n_j \rangle$:

$$v_{\langle n_i , n_j \rangle} = subTime\left(\langle n_i , n_j \rangle\right). \quad (7)$$

If $v_{\langle n_i , n_j \rangle}$ meets the threshold requirements, $\langle n_i , n_j \rangle$ can be regarded as an associated newsgroup. And the consumer news recommendation list can be obtained according to these associated newsgroups.

The above of all is the specific implementation steps of the personalized recommendation algorithm based on consumers' click behavior. Next, some experiments will be done to verify this algorithm.

IV. Experiment and Analysis

A. Data Set

In this paper, data set is obtained from the domestic well-known financial website-caixin.com. The data set consists of two parts, the first part is train data and it contains all the news browsing history of 10000 consumers in March 2014. Every history includes five parts, namely consumer ID, news ID, and browsing time, title of news and content of news. The number of history is 116225. Another part is test data and it contains the last browsing history of 10000 consumers. So, the number of history is 10000. In this experiment, the train data is used to predict the test data with the aid of personalized recommendation algorithm based on consumers' click behavior.

B. Performance Measure

Let F-measure evaluate the above algorithm. F-measure is the harmonic mean of precision and recall. The precision is the fraction of retrieved instances that are relevant, and the recall is the fraction of relevant instances that are retrieved. Both precision and recall are based on a measure of relevance. F-measure can be calculated by the following formula:

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}. \quad (8)$$

Let $hit\left(U_i\right)$ represent the number of correct result recommended to U_i, the $hit\left(U_i\right)$ can only take 1 or 0 in this experiment, and let $L\left(U_i\right)$ represent the number of news recommended to U_i. Let $T\left(U_i\right)$ be used to represent the number of last browsing history of U_i, and it can only take 1 in this experiment.

$$precision = \frac{\sum_{i=1}^{N} hit\left(U_i\right)}{\sum_{i=1}^{N} L\left(U_i\right)}. \quad (9)$$

$$recall = \frac{\sum_{i=1}^{N} hit\left(U_i\right)}{\sum_{i=1}^{N} T\left(U_i\right)}. \quad (10)$$

C. *Experiment and Analysis*

According to the above algorithm, let T_l be set with 1 and T_h be set with 1000. And the best recommendation result is obtained through changing value of confidence. The following charts show the trends of precision, recall, F-measure and operation efficiency.

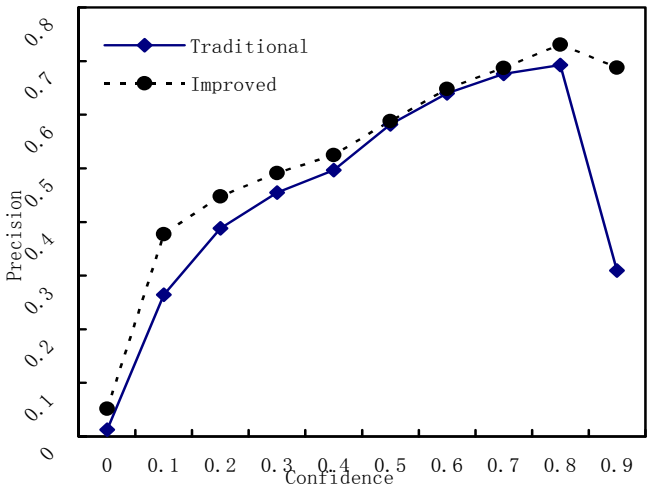

Figure 1. The trend of precision

From fig. 1, the precision of recommendation has improved. The limit of browsing time difference decreases negative impact of low association rules on the recommendation result. But, the number of recommendation will also be reduced because of decrease of association rules. And the following figure shows this change.

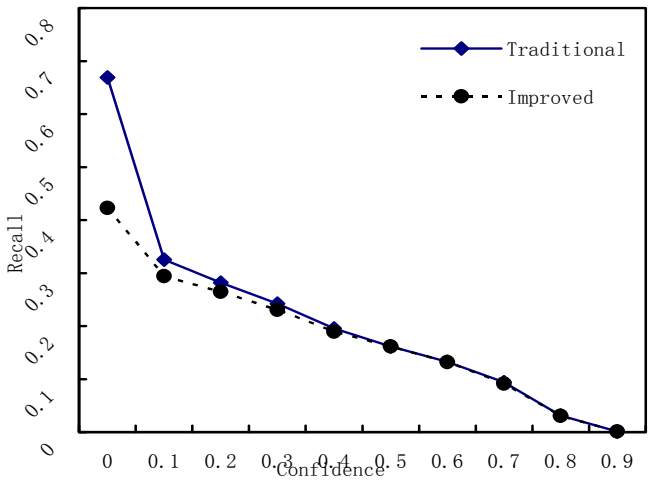

Figure 2. The trend of recall

In fig. 2, the recall of recommendation has dropped. The decrease of low association rules will also decrease the number of correct recommendation, because low association rules can also produce recommendation list with correct results. In fact, the improved recommendation algorithm just provides a method to remove result sets with low accuracy from the recommendation result which is produced by traditional recommendation algorithm.

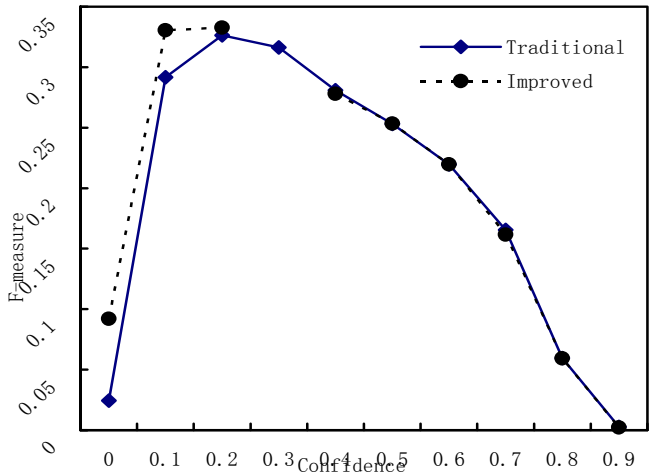

Figure 3. The trend of F-measure

From fig. 3, the improved recommendation algorithm is better than the traditional recommendation algorithm under different confidence levels overall. And the highest value of F-measure is 0.33262. Although the improved recommendation algorithm is lower in the recall, it has a better performance than the traditional algorithm on the whole.

TABLE II. THE TREND OF OPERATION EFFICIENCY

Confidence	***0.0***	***0.1***	***0.2***	***0.3***	***0.4***
Traditional	***76***	***112***	***111***	***130***	***124***
Improved	***63***	***92***	***91***	***95***	***78***
Confidence	***0.5***	***0.6***	***0.7***	***0.8***	***0.9***
Traditional	***122***	***136***	***117***	***116***	***114***
Improved	***84***	***97***	***95***	***92***	***96***

The data in table is the operation time(s)

In table II, the operation time of recommendation algorithm is presented. Because the data set is relatively large, the operation time is long. Overall, the improved recommendation algorithm has obvious advantages than the traditional recommendation algorithm. The decrease of low association rules make program do not have to spend much time on calculating recommendation based these rules, so the efficiency is promoted.

V. CONCLUSION

The personalized recommendation algorithm based on consumers' behavior obtains the similarity of consumers through finding potential association between the news browsing sequences of consumers. But traditional recommendation algorithm ignores the importance of browsing

time. In this paper, an improved recommendation algorithm is proposed. The factor of browsing time is added into the construction of association rules. The experiments show that the proposed recommendation algorithm is better than the traditional recommendation algorithm.

However, the proposed algorithm pays more attention to the efficiency but ignores the accuracy. And the algorithm just takes advantage of one attribute of consumers' behavior, namely browsing time. So the recommendation result of proposed algorithm gets a little higher than the recommendation result of traditional algorithm. In the following research, more attributes will be added into experiment, such as duration of behavior, and more works will be focused on the improvement of accuracy.

Now, there are some ways worth getting a try. The change of consumers' behavior can reflect the change of consumers' interest. So recommendation can be obtained from the change. In addition, accidental behavior may contain some potential consumers' information. Such behavior is abundant in the Internet, so it is useful for recommendation to mine the potential information.

ACKNOWLEDGEMENT

This paper is partly supported by the National Bureau of Statistics Project (Project number: 2011 LY074), "The comprehensive reform project of computer science and technology, department of science and Engineering" , National Youth Natural Science Foundation of China (Grant No. 61305100), "Guangzhou Research Institute of Communication University of China Common Construction Project, Sunflower – the Aging Intelligent Community", "The comprehensive reform project of computer science and technology (Project number: ZL140103)" and Engineering Project of Communication University of China (Pr oject number: 3132015XNG1504).

REFERENCES

[1] J. G. Liu, T. Zhou, and B. H. Wang, "Research progress of personalized recommendation system," Progress in natural science, vol. 19, pp. 1-15, Jan. 2009.

[2] C. S. Cui and Q. Z. Wu, "Research on content-based recommendation based on Vague sets", Application Research of Computer, vol. 27, pp. 2109-2111, Jun. 2010

[3] B. Yang and P. F. Zhao, "Survey of recommendation algorithm", Journal of Shanxi University (Nat. Sci. Ed.), vol. 34, pp. 337-350, March 2011.

[4] Y. J. Leng, Q. Lu and C. Y. Liang, "Survey of recommendation based on collaborative filtering", PR & AI, vol. 27, pp. 720-734, Aug. 2014.

[5] D. T. Sun, T. He and F. H. Zhang, "Survey of cold-start problems in collaborative filtering recommender system", Computer and modernization, vol. 201, pp. 59-63, May. 2012.

[6] J. M. Chen, Y. Tang, J. G. Li and Y. B. Cai, "Survey of personalized recommendation algorithms", Journal of South China Normal University (Natural Science Edition), vol. 46, pp. 8-15, Sep. 2014.

[7] C. Zhang, G. Chen, and H. M. Wang, "Recommendation model based on blending recommendation technology", Computer Engineering, vol. 36, pp. 248-250+253, Nov. 2010

[8] Nakagawa, Miki, and B. Mobasher. "Impact of site characteristics on recommendation models based on association rules and sequential patterns", Proceedings of the IJCAI. Vol. 3. 2003.

[9] Agrawal, Rakesh, T. Imieliński, and A. Swami. "Mining association rules between sets of items in large databases", ACM SIGMOD Record. Vol. 22. No. 2. ACM, 1993.

[10] Agrawal, Rakesh, and R. Srikant. "Fast algorithms for mining association rules", Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215. 1994.

[11] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. "Effective personalization based on association rule discovery from web usage data", Proceedings of the 3rd international workshop on Web information and data management. ACM, 2001.

[12] Nasraoui, Olfa, and R. Krishnapuram. "One step evolutionary mining of context sensitive associations and web navigation patterns", in SIAM conference on Data Mining. 2002.

[13] Awad, Mamoun, L. Khan, and B. Thuraisingham. "Predicting WWW surfing using multiple evidence combination", The VLDB Journal—The International Journal on Very Large Data Bases 17.3 (2008): 401-417.

[14] M. Awad and L. Khan. "Web navigation prediction prediction using muitiple evidence combination and domain knowledge". IEEE Trans. Syst., Man, Cybern. A, Syst., Hummans, vol. 37(6), pp. 1054-1062, Nov. 2007.

[15] J. Liu, P. Dolan, and E. Pedersen. "Personalized news recommendation based on click behavior". In Proc. Of 15th Int. Conf. on IUI, 2010, pp. 31-40.

[16] M. T. Hassan, K. N. Junejo, and A. Karim. "Learning and predicting key web navigation patterns using Bayesian model". In Proc. Int. Conf. Comput. Sci. Appl. II, Seoul, Korea, 2009, pp. 877-887.

[17] Speretta, Mirco, and S. Gauch. "Personalized search based on user search histories", Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on. IEEE, 2005.

[18] Tan, Pang-Ning, M. Steinbach, and V. Kumar. "Introduction to data mining". Vol. 1. Boston: Pearson Addison Wesley, 2006.

[19] J. Wang, and X. C. Tang, "Personalized recommendation algorithm research based on content in social network", Application Research of Computer, vol. 28, pp. 1248-1250, Apr. 2011.