International Journal of the Physical Sciences Vol. 6(14), pp. 3518-3523, 18 July, 2011
Available online at http://www.academicjournals.org/IJPS
DOI: 10.5897/IJPS11.485
ISSN 1992 - 1950 ©2011 Academic Journals

Full Length Research Paper

Hybrid filtering model based on particle swarm optimization and genetic algorithm

ZHU Zhenfang[1]*, LIU Peiyu[2], ZHENG Yan[1], ZHAO Jing[1], LI Shaohui[2] and WANG Jinlong[2]

[1]School of Information Science and Engineering, Shandong Normal University, Ji'Nan 250014, China.
[2]Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, Ji'Nan 250014, China.

Accepted 11 May, 2011

With the rapid growth of network information, information filtering technology is more widely used. This paper discusses the content-based filtering and collaborative filtering, and proposes a hybrid filtering model with these two methods in order to overcome their own shortages. In this hybrid filtering method, genetic algorithm is used to generate initial profiles on server-side, and particle swarm optimization is used to update the profiles with the information from users. This approach is feasible from the theoretical analysis and the experiment in Chinese data set.

Key words: Information filtering, content-based filtering, collaborative filtering, particle swarm optimization, genetic optimization.

INTRODUCTION

With the development of information technology, information world provides plenty of network information to computer users, however, people are also inevitably exposed to a lot of spam while they enjoy the convenience of information. Hence, the network information filtering arises at this historic moment.

THE PROBLEM STATEMENT

The information filtering (Belkin and Croft, 2009) is an automatic process, which could get the useful information and remove useless information from the large-scale dynamic information stream according to user's information needs.

Text filtering is an important part of information filtering, TREC-9 gives the definition of text filtering (Robertson and Hull, 2001) according to the given user needs, text filtering system establishes a filter template, which could select relevant texts automatically from text stream. Using this profile, text filtering system automatically accepts or rejects the texts and correct filter template adaptively according to feedback information.

In a certain extent, text information filtering can be regarded as a kind of binary text classification, the text to be filtered will be mapped to a legitimate or illegal text collection. This process can be expressed with mathematical language as follows:

For each $< d_i, c_i > \in D \times C$ (Where D is the document set to be filtered, d_i is one document of D, C is the set of category, and it has two value called filtering text collection and normal text collection, c_i is a value of C. To determine the value of document C, if the value is TRUE (T), the text d_i is a number c_i, otherwise, it does not belongs to c_i. So, the process of information filtering is described by constructed function α:

$$D \times C \Rightarrow \{T, F\}$$

THE OBJECTIVE OF STUDY

In a broad sense, information filtering includes text

*Corresponding author. E-mail: zhuzhfyt@163.com.

Table 1. User-Item matrix.

	Item 1	Item 2	Item 3	Item 4
User 1	?	2	3	4
User 2	3	?	2	2
User 3	3	4	1	?
User 4	1	2	4	3

filtering, audio filtering, photograph filtering, and video filtering, and so on (Huang et al., 2003). But in a narrow sense, it means text information filtering in particular (Bedi et al., 2009). Meanwhile, most of the information spreads on internet is text now, so the information filtering in this paper is text information filtering, especially Chinese text information filtering.

Text filtering is a process, which could search out texts from the large-scale text data stream to the specific user's needs. The main methods used in information filtering is content-based information filtering and the collaborative filtering, these two methods have their own advantages and disadvantages (Benczur et al., 2009).

This paper discusses the content-based filtering and collaborative filtering, and proposes a hybrid filtering model with these two methods in order to overcome their own shortages. In this hybrid filtering method, genetic algorithm is used to generate initial profiles on server-side, and particle swarm optimization is used to update the profiles with the information from users. This approach is feasible from the theoretical analysis and the experiment in Chinese data set.

THE CRITICAL REVIEW OF THIS RESEARCH

Content-based filtering

Content-based filtering is a primary method of information filtering. In the content-based information filtering, the filtering system generates document vectors from training documents and documents which will be filtered, using word segmentation, feature item weighting, dimensions reducing and other preprocessing methods, and then gets content-based profile by learning from training documents vectors, at last, disposes the document to be filtered by the similarity of document vector between the profile and the document to be filtered. On the basis of this, filtering system could be maintain by modifying these profiles from users' feedbacks and improve the information filtering quality (Lin et al., 2000).

Content-based filtering technique filters information according to the similarity of network documents and filtering profile, during the filtering process, each user uses the relevant profile to filter information without thinking about other users' filtering results and methods. Though content-based filtering is simple and effective, it also has problems:

(1) It could not differentiate the filtering quality in the same category: The categories of profile are determined by training documents because the filtering profile generates from training documents. For example, we use academic literatures of sports theory to train some kind of sport categories, but in the real filtering, there are application literatures or news report, in that way, there would be problems during the filtering process.

(2) Few alters after generating profile: Filtering profile are generated by learning, once generated, could not be altered fundamentally, even the dynamic adjusting strategies are used, so filtering profile could not find the new relevant information of the same category.

(3) The algorithms affect filtering effectively: Different learning method is applied to different training documents, so during the training and filtering process, the filtering results would be very inaccurate if we use the inappropriate content-based filtering method.

Techniques of collaborative filtering

Collaborative filtering is also widely used (Su et al., 2009; Lathia et al., 2009), it could give the category of documents to be filtered by the collaboration of users. And its starting point is the relationship of users, because the interest of each person is not isolated, but should be in a group.

In the collaborative filtering, the filtering system analyses user's interests, then find users which has the same or similar interests to the active user, and combine the evaluations of these users to that information, at last, give the prediction of the active user. The principle of collaborative filtering could usually use the User-Item matrix, as shown in Table 1.

In the matrix, the values which we could see is the interest evaluation to the relevant information, and the unknown values would be given by the system. Collaborative filtering is a process to predict unknown values by the known values.

Compared with content-based filtering, collaborative filtering has many merits, such as, it could filter the information which is hard to distinguish by content or complicated to express, and it has the capability to

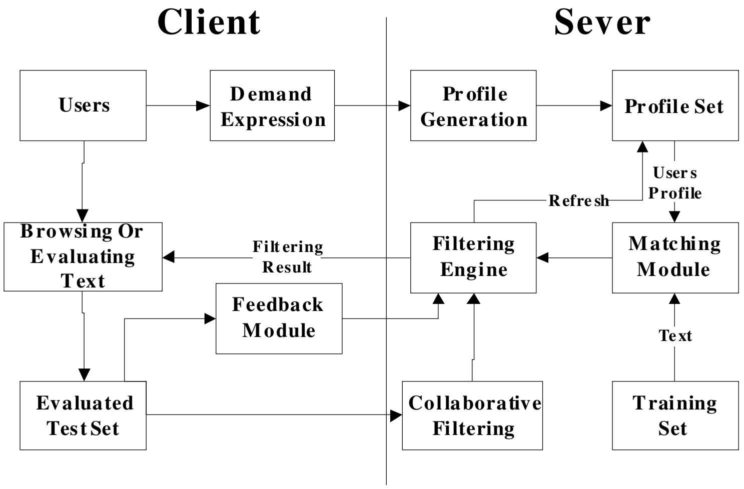

Figure 1. Structure of hybrid filtering model.

recommend. But it also has disadvantages:

(1) Lack of initial data: There is few evaluated information at first because of few number of users. As a result, the few information which is evaluated or not evaluated could not be classified and filtered.

(2) Data sparsity: As the number of information item is much larger than the number of user, it could be accepted that few users would like to evaluate the information they have browsed, even if they give the evaluation, the number of evaluation is small, so it makes the user-item matrix sparse, there are few evaluation levels in the prediction computing, so it is very hard to find similar users and filter accurately.

(3) Big computing amount later: The users-based prediction algorithm computes users' similarities from all the system users, the compute amount would be bigger and the efficiency of the system would be lower with the increase of system users and information source.

FEASIBILITY OF HYBRID FILTERING

Necessity

For the first and second problems of content-based filtering, the collaborative filtering could solve them easily, because the collaborative filtering could take full advantage of users' evaluations. In the collaborative filtering, users give high or low evaluation after they read different information, and in the next collaborative filtering, the system could provide a better recommenddation to the active user by the interest evaluation from the same or similar users, by doing this, it makes the information on one topic to be differentiated easily, and it also could recommend some new information at the same time. For the third problem of content-based filtering, if combined with collaborative filtering, the accuracy of hybrid system would be improved with accuracy of collaborative filtering improving when the evaluation levels increase.

The first and second problems of collaborative filtering would be solved if combined with content-based filtering. Content-based filtering could filter information by similarity between the information source and users' interests, it could recommend the high similarity information to users without thinking about the evaluation, so the information which have not been read or evaluated could be filtered and recommended to users. If combined with content-based filtering, the users could give evaluations to the filtering results, and increase the number of evaluation levels, so the content-based filtering would work without the influence of initial evaluation levels problem and the problem of data sparsity, and they could be solved very well.

Framework of hybrid filtering model

The content-based and collaborative filtering techniques have been used in real systems at home and abroad, and the performance of these systems improves a lot (Ilyas et al., 2010). During the combination of these two filtering techniques, the main problem is how to combine. In this paper, the content-based filtering profiles are generated by genetic algorithm (Algarni et al., 2009), and the coordination between clients, client and sever is optimized by PSO (Claypool et al., 1999). The basic framework is shown in Figure 1. User model generation module generates interest models by analyzing different

Table 2. Training set distribution.

Category	Violence	Pornography	Computer	Environment
Number	276	192	1358	1218
Category	Agriculture	Economic	Politics	Sports
Number	1022	1601	1026	1254

Table 3. Contingency table.

	Text number classified correctly	Text number classified falsely
Judge to the category	a	b
Judge to other categories	C	d

users' information needs. Matching module is in charge of content filtering, browsing or evaluating module allows users to browse filtering results in a visual way and give the interest evaluation feedbacks. The evaluated test set saves texts evaluated by users, which could be used in feedback module to refresh new users' interest model and could also be used in collaborative module to collaborative filtering recommend, filtering engine is scheduling core in system, it sends the filtering results of matching module or collaborative module to users initiatively, and could refresh users' interest models by users' feedbacks (Taishan, 2010).

EXPERIMENT DATA AND ENVIRONMENT

Training sets

The training set comes from the corpus sorted by LI Rong-lu (Natural Language Processing Group, the International Database Centers, Department of Computer Information and Technology, Fudan University), in this corpus, there are 20 categories and 9804 texts. In these texts, 11 categories have less than 100 texts, and 6 categories have more than 1000 texts, such as: Computer, environment, and economy and so on. Because this classifier would be used to filter spam, the project team collected two categories called pornography and violence. The distribution of training set is as shown in Table 2.

Test sets

The test sets which have a total of 902 texts, consists of two parts, one is the categories which have no more than 100 texts of corpus sorted by LI Rong-lu, and the other comes from the training sets, and each category randomly selects 50 texts.

Development environment

The development environment comprises the Founder PC, Processor: Intel (R) Core (TM) Duo CPU E7200 @ 2.53HZ, Memory: 1G, Programming environment: Visual Studio 2005 and Development Language: C#.

Performance validation

The commonest evolution methods (Miao and ZH, 2007) of classification and filtering are precision (p) and recall (r), for each category o, researchers use contingency table to calculate precision (p) and recall (r). Table 3 gives an example of contingency table.

On the basis of contingency table, we could define precision (p) and recall (r) as follows:

$$p = \frac{a}{a+b} \qquad r = \frac{a}{a+c}$$

Contingency table (Table 3) could only evaluate single category, and if we want to evaluate the classifier, we should use another parameter named Micro-averaging:

$$\bar{r} = \frac{\sum_{1}^{|c|} r_c}{|c|} \qquad \bar{p} = \frac{\sum_{1}^{|c|} p_c}{|c|}$$

PROGRAM OF HYBRID FILTERING

From Figure 1 we could see that, in the hybrid filtering model, we could divide the process into two parts, content-based filtering on the servers and collaborative filtering by the cooperation between the clients and the servers.

Content-based filtering based on genetic algorithm

Network information filtering model based on genetic algorithm has been briefly addressed in the literature (Zhu et al., 2010, Zhu and Liu, 2010).

Collaborative filtering based on PSO

Particle swarm optimization (Gao et al., 2009) is a random search optimization algorithm, which is population basic and self-adapting, invented by Eberhart and kennedy (1995). Its basic ideas stem from the simulation of birds flock's looking for food, and it also take advantage of biosystem modeling coined by Frank Heppner. Now, it has been widely used in function optimization, constraint optimization,

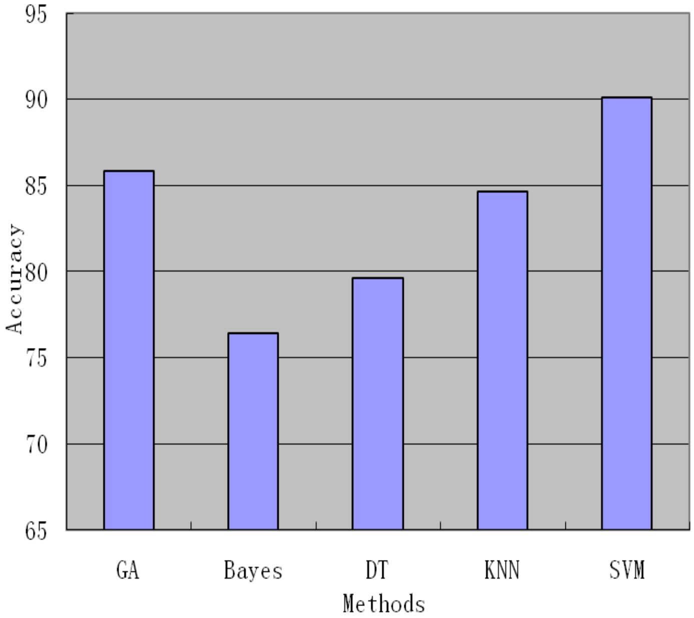

Figure 2. Comparison of average accuracy.

minimax optimization, and so on. And this method could also be combined with various methods for finite difference equation (Yi et al., 2011). In this paper, the PSO is introduced in information filtering to improve information filtering effect by optimization of coordination between the clients and the servers.

In the hybrid filtering model, the collaborative filtering part would undertake updating the profile by data exchange of the clients and servers, and the ultimate goal is enhancing the filtering results, including recall, precision, F1 value, and so on.

In PSO, the information filtering is taken as a multi-objective optimization question of recall, precision and F1 value. In this problem, we take the initial profile and the initial training sets for the cognitive factor and take the data from clients for the social factor, and take the recall, precision and F1 value as the driving force for PSO.

Main process of hybrid filtering model

As the main points of content-based information filtering and collaborative filtering are shown, the main process of hybrid filtering model is as follows:

Step 1: Generate the filtering profile based on the genetic algorithm on server.

Step 2: Send the profile to the clients which are considered as particles.

Step 3: The client filter information and submit the result to the server and the other clients, at the same time, the server and the clients also keep records of the online behavior which is considered as the social factor of PSO, such as browsing history.

Step 4: Take the initial profile and the initial training sets for the cognitive factor, update the profile combined with the social factor we have talked about.

Step 5: Ascertain if the profile need to update by judging the new target value, and if it needs to be updated, the sever would send the update to the client.

RESULTS AND DISCUSSION

From the calculation, the average accuracy of data is $\bar{p}$ = 85.810, we compared our result with several basic methods which is shown in Figure 2. In these methods (Su et al., 2006), the method is better than Naive Bayes method, tree method and nearest neighbor classification method. The comparism is shown in Figure 2.

In the information filtering test, we divided the test data into two categories, namely: legal and illegal documents. In the test set, the illegal documents are composed of the test data of illicit sex and violence, and the legal documents are composed of legitimate documents randomly selected from six other categories, experimental results show that the accuracy of illegal document filtering has reached 97.67% and it is better than that in the literature achieved by Zhu and Liu (2010).

CONCLUSIONS

With the spread of internet and the growth of information needs, content-based information filtering and the collaborative filtering, have been widely used in the information recommendation of electronic commerce and e-library. As regards the problems encountered with these two methods, this paper gave an improved filtering model, which composted these two methods. We could focus on improvement of the filtering model based on genetic algorithm, methods of information collection, improvement of PSO, and so on.

ACKNOWLEDGMENTS

This work is supported by National Nature Science Foundation of China (60873247), Nature Science Foundation of Shandong Province (ZR2009GZ007), Independent Innovation of High and New technology Special Project of Shandong Province (2008ZZ28). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

Algarni A, Li YF, Xu Y, Lau Raymond YK (2009). An effective model of using negative relevance feedback for information filtering. Proceeding of the 18th ACM Conference on Information and Knowledge Manag.2-6 November 2009, Asia World-Expo., pp. 121-154.

Benczur AA, Miklós E, Julien M, David S (2009). Web Spam Challenge Proposal for Filtering in Archives. Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web ACM, Madrid,Spain.1-2.http://portal.acm.org/citation.cfm?id=1531928.

Bedi P, Sharma R, Kaur H (2009). Recommender system based on collaborative behavior of ants. Int. Artif. Intell., 2: 40-55.

Belkin NJ, Croft WB (1992). Information filtering and information retrieval: Two sides of the same coin? Commun. ACM., 35(12):29-38.

Claypool M, Gokhale A, Miranda T, Murnikov P, Netes D, Sartin M (1999). Combining content-based and collaborative filters in an online newspaper. Proceedings of the ACM SIGIR Workshop on Recommender Systems-implementation and Evaluation. Aug. 9-19. ACM Press, USA.1-8.

Gao XY, Sun LQ, Sun DS (2009). An enhanced particle swarm optimization algorithm. Inform. Technol. J., 8: 1263-1268.

Huang XJ, Xia YJ, Wu LD (2003). A text filtering system based on vector space model. J. Software, 14(3): 435-442.

Ilyas MZ, Samad SA, Hussain A, Ishak KA (2010). Improving speaker verification in noisy environments using adaptive filtering and hybrid classification technique. Inform. Technol. J., 9: 107-115.

Lathia N, Hailes S, Capra L (2009). Evaluating collaborative filtering over time. Proceedings of SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.

Lin HF, Li YL, Yao TS (2000). An information influence mechanism for Chinese text filtering. J. Comput. Res. Dev., 37 (4):470-4763

Miao DQ, Wei ZH (2007). Principle and Application Chinese Text Information Processing Beijing: Tsinghua University Press, 228-230.

Robertson S, Hull DA (2001). The TREC-9 filtering track final report (2001). Proceedings of the 9th Text Retrieval Conference (TREC-9).Gaithersburg: NIST Special Publication, pp. 25-40.

Su JS, Zhang BF, Xu X (2006).Advances in machine learning based text categorization. J. Software., 17(9):1848-1859

Su XY, Khoshgoftaar TM (2009). A survey of collaborative filtering techniques. Adv. Artificial Intell., pp. 1-19.

Taishan Y (2010). An improved adaptive genetic algorithm based on human reproduction mode for solving the knapsack problem. Inform. Technol. J., 9: 974-978.

Yi Y, Yu ZQ, Ye W (2011). Modified particle swarm optimization and genetic algorithm based adaptive resources allocation algorithm for multiuser orthogonal frequency division multiplexing syst. Inform. Technol. J., 10: 955-964.

Zhu ZF, Liu PY (2010). Feasibility research of text information filtering based on genetic algorithm, Sci. Res. Essays, 5(22): 3405-3410.

Zhu ZF, Liu PY, Zhao LN, Lv TX (2010). Research of feature weights adjustment based on Semantic paragraphs matching. ICIC Express Lett., 4(2): 559-564.