Personalized Web Information Recommendation Algorithm Based on Support Vector Machine

Yu Bo[1] and Qi Luo [2]
[1] School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[2] School of Electrical and Information Engineering, Wuhan Institute of Technology, Wuhan 430074, China
E-mail boboyu615@gmail.com

Abstract

With the explosion of Web information, how to immediately and exactly find the needed information for each user has become a tough problem. To meet the personalized needs of users in information service, a new personalized recommendation algorithm based on support vector machine was proposed in the paper. First, user profile was organized hierarchically into field information and atomic information needs, considering similar information needs in the group users. Support vector machine was adopted for collaborative recommendation in classification mode, and then Vector Space Model was used for content-based recommendation according to atomic information needs. The algorithm had overcome the demerit of using collaborative or content-based recommendation solely, which improved the precision and recall in a large degree. It also fits for large scale group recommendation. The algorithm was also used in personalized information recommendation service system. The system could support information recommendation better. The results manifested that the algorithm was effective.

1. Introduction

With the explosion of Web information, how to immediately and exactly find the needed information for each user has become a tough problem [1]. Although traditional technologies of search engine meet some demands of users, they cannot fulfill the personalized requirements of users in various backgrounds, with diverse intention and at different time. So, many scholars have carried on a great deal of researches on personalized recommendation algorithms, such as collaborative recommendation algorithm and content-based recommendation algorithm [2] [3]. Although collaborative recommendation algorithm is able to mine out potential interests for users, it has some disadvantages such as sparseness, cold start and special users. Similarly, the content-based recommendation algorithm also has some problems such as incomplete information that mined out, limited content of the recommendation, and lack of user feedback.

According to this, a new personalized recommendation algorithm based on support vector machine is proposed in the paper. It has improved the traditional algorithm. First, user profile is organized hierarchically into field information and atomic information needs, considering similar information needs in the group users. Support vector machine is adopted for collaborative recommendation in classification mode to ensure the recall, and then Vector Space Model is used for content-based recommend according to atomic information needs, which ensures the precision. The algorithm is also used in personalized Web information recommendation service system. The results manifest that it can support Web information recommendation better.

2. User Profile

Supposed that some information fields' space S_d constructs global information space $S_{global} = S_{d1} \bigcup S_{d2} \bigcup \cdots S_{dn}$, space of personal information need is the subset of space of some field information $DIS = DIS_{i1} \bigcup DIS_{i2} \bigcup \cdots \bigcup DIS_{in}$. DIS is the subset of GIS, the relations among them as follows [4]:

$$PIS \subseteq DIS \subseteq GIS$$

In fact, in large scale group users, users' information need has a big superposition. $PIS_{i1} \bigcap PIS_{i2} \bigcap \cdots PIS_{in} \neq \varnothing$

0-7695-3006-0/07 $25.00 © 2007 IEEE
DOI 10.1109/IPC.2007.107

Usually, a user may interest in some fields. Organizing users' information need hierarchically is not only satisfied with users' actual need, but also the system can provide recommendation service conveniently. Thus, the paper uses this method to organize users' information need, the structure is Fig .1

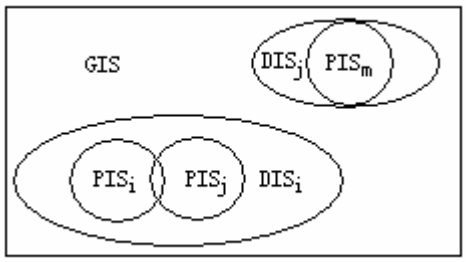

Figure1 Relationship of user information

Space of users' information need could be seen as a U= {UID, PIS, Period, Freq}

UID=inter UID || Exter UID

PIS=GIS || DIS || PIS

DIS=DIS_1 || DIS_2 || $DIS_3 \cdots$ || DIS_k

$PIN = PIS_1 \,||\, PIS_2 \cdots ||\, PIS_m$

$PIS_i = < I_i, w_i >$

$Period = StartTime \,||\, EndTime$

Freq=Hour || Day || Week || Month

$< I_i, w_i >$ Separately represents item and weighting of users' actual need. StrartTime and End time separately represent the time of system service

Form the structure, a model of users' information need is a tree. Root node is user UID, the middle node is corresponding to the catalog of field information. Leafage node saves actual need of current field information PIS_i . PIS_i Is not separable, which is called atomic information need. The tree is Fig. 2

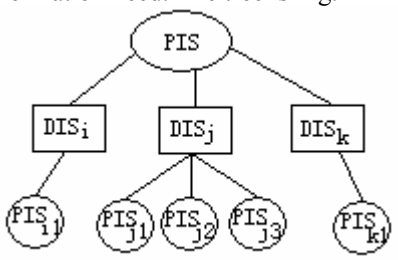

Figure2 Hierarchical information need

3. Personalized recommendation algorithm based on Support Vector Machine

3.1. Recommendation algorithm

Field need information is used by classification recommendation, and then users' actual atom information need is used by content-based recommendation. There are two phases:

(1) First filtering. Users' field information need is classified by SVM, which ensure the recall. This strategy considers user recessive need.

(2) Second filtering. Basing on first filtering, Second filtering realizes user atom information need recommendation by VSM, which ensure the precision.

3.2. Field Information Needs classification

Support vector machine for classification can be seen as the application of perceptron [5]. When classification problem is linearly separable, a hyperplane that makes two categories of separable data more close to the plane is set up（usually the plane is called optimal separation hyperplane**.** Regarding to nonlinear problem, original data is mapped from a low dimensional space to the new data sets of a higher dimensional space (feature space) through a nonlinear mapping (nuclear function).New data sets are linearly separable in feature space, thus the classification in higher dimensional space is completed.

Basing on it, we can put the question of nonlinear separable and linear separable into a unified formula. Regarding to the sample space $\Omega = \{(x_i, y_i) \,||\, i = 1, 2, \cdots, N\} \subset R^n \times \{-1, 1\}$ and the function $\{\Theta(x_i), \Theta(x_j)\} = K(x_i, x_j)$ Standard support vector machine can represent as follows:

$$\min Q(w, \varepsilon) = \frac{1}{2} \| w \|^2 + C \sum_{i=2}^{l} \varepsilon_i \qquad (1)$$

$$s.t. y_i [\{w \bullet \Theta(x_i) + b] - 1 + \varepsilon_i \geq 0 \,, \quad \varepsilon_i \geq 0, i = 1, \cdots, l \qquad (2)$$

(a). When $K(x_i, x_j)$ is linear transform, especially when $K(x_i, x_j)$ is linear invariant mapping, and C=0，$\forall \varepsilon_i = 0$, (1) (2) is correspond to linear separable condition.

(b). $K(x_i, x_j)$ is nonlinear mapping that transforms Ω to a higher dimensional space H, and $K(x_i, x_j)$ satisfies Mercer theorem. If $K(x_i, x_j)$ is

Kernel Function, (1) (2) is correspond to nonlinear separable condition. C>0 is a constant, which control the punishment degree of misclassification. Loss function $\sum_{i=1}^{l}\varepsilon_i$ is an upper bound of misclassification. In fact, $\sum_{i=1}^{l}\varepsilon_i$ can represent as

$$F_\sigma(\varepsilon)\sum_{i=1}^{l}\varepsilon_i^\sigma \quad (3)$$

σ =1 is corresponding to support vector machine of one time loss function, σ =2 is correspond to support vector machine of twice time loss function.
(1)－(2) can be summarized to solve the problem of quadratic programming.

$$Min\ \ W(a) = -\sum_{i=1}^{l}a_i + \frac{1}{2}\sum_{i,j=1}^{l}a_i a_j y_i y_j K(x_i \bullet x_j) \quad (4)$$

$$s.t:\ 0 \le a_i \le C, i = 1,\cdots,l;\ \sum_{i=1}^{l}a_i y_i = 0; \quad (5)$$

That is corresponding to decision functions of the broadest and optimal classification

$$f(x) = \mathrm{sgn}(\sum_{S.V.} y_i a_i K(x_i \bullet x) + b) \quad (6)$$

In fact, a classifier is a judge function, which makes variable D in definition domain divide to inconsistent range subspaces $C = \{C_1, C_2, C_3, \cdots, C_n\}$ though certain principles, $C_i \bigcap C_j = \varnothing$ the input variables is the most approach to another defined category according to some testing degree σ .

$$f_\sigma = (x_0) = \arg\max_{c_{i\in C}}(c_i, x_0) \quad (7)$$

3.3. Atom Information Needs Recommendation

Salton method is used to compute the similarity. In Salton method, each text is represented as an n dimensional vectors W= ($w_1, w_2, w_3, \cdots, w_n$), component w_i is corresponding to the weigh of feature in this text. The measurement of w_i as follows:

$$w_i = \frac{tf_i \times \log(N/n_i)}{\sqrt{\sum_j (tf_j \times \log(N/n))^2}} \quad (8)$$

While, tf_i is occurrence times of feature in assigned texts, N is the total number of text in set of texts, n_i is occurrence times of feature.

The similarity s is computed according to the cosine value between user need vectors and every text.

$$S = COS(w_u, w_d) = \frac{\sum_{i=1}^{n} w_{ui} \bullet w_{di}}{\sqrt{\sum_{i=1}^{n}(w_{ui})^2 \bullet \sum_{i=1}^{n}(w_{di})^2}} \quad (9)$$

For recommendation results, real information needs and potential users information needs should be considered. So, the recommendation results R are divided into positive region sets R^+ and negative region sets R^- , which is corresponding to user real needs and potential needs, and thus system recommendation decision as follows:

$$R = \begin{cases} R^+, s \ge s_0^+ \\ R^-, s_0^- \le s \le s_0^+ \\ discard, s < s_0^- \end{cases} \quad (10)$$

s_0^+, s_0^1 are represented separately as recommendation of positive region and negative region.

4. Conclusion

On the foundation of research, the author constructs a personalized information recommendation system website. Evaluation metrics as follows [6]:

$$\mathrm{Pr}ecision = \frac{number\ of\ \inf ormation\ correctly\ \ filtered}{number\ of\ \inf ormation\ \ filtered} \quad (11)$$

$$\mathrm{Re}call = \frac{number\ of\ \inf ormation\ correctly\ \ filtered}{number\ of\ \inf ormation} \quad (12)$$

$$F_\beta = \frac{(\beta^2+1)\times Precision \times \mathrm{Re}call}{Precision + \mathrm{Re}call} \quad (13)$$

In order to obtain the contrast experimental result, the SVM classification algorithm, content-based recommendation algorithm based on VSM and a personalized recommendation algorithm are separately used in the module of personalized recommendation. The experimental data comes from web data, 9888 texts are selected as a test sets, 2030 texts are selected as a feedback evaluation sets. 6 categories of texts are

chose to do the experiment. In the recommended classification phase, BSVM classifier based on LIBSVM is used as classification. In content-based recommendation phase, similarity range satisfies $s_0^+ = 2s_0^-$ .Supposed s_0^+ =0.2, three types of texts are recommended. The quantity of each category texts is five. X axis represents recall, Y axis represents precision. The experimental results are Fig.3.

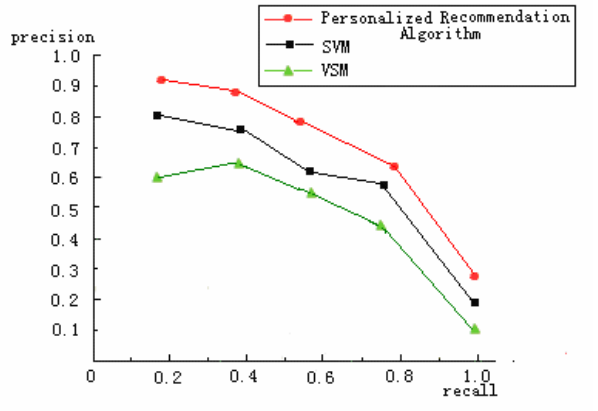

Figure3 the performance of personalized recommendation algorithm, SVM and VSM

From Fig.3, the precision of personalized recommendation algorithm based on support vector machine is higher than separate classification algorithm. The recall of personalized recommendation algorithm based on support vector machine is more effective than separate content-based recommendation algorithm.

In summary, a personalized web information recommendation algorithm based on support vector machine is proposed in the paper. The algorithm is also used in personalized information recommendation service. The results manifest that the algorithm is effective through testing in personalized information recommendation service system.

References

[1]Yifeng Xu. The Study on the Application of Rough set Theory in the Web Information Filtering. Computer Systems Aplications.2007.03.

[2]Zhang Bingqi.A Collaborative Filtering Recommendation Algorithm Based on Domain Knowledge.Computer Engineering, Beijing, vol.31, 2005, pp.29-33.

[3]Deng Ailin.Collaborative Filtering Recommendation Algorithm Based on Item Clustering. Mini-Micro System.vol25, 2005, pp.1665-1668.

[4]Qi Luo.Research on personalized service system in E-supermarket by using adaptive recommendation algorithm. Journal of Communication.Vol.27, No.11, 2006.pp.183-187.

[5]Vapnik V.The nature of statistical learning theory.Tsinghua University Press.Beijing, 2000, pp.162-163.

[6] Li Dun and Cao Yuanda.A New Weighted Text Filtering Method.International Conference on Natural Language Processing and Knowledge Engineering.wuhan, 2005 pp695-698.