Research on WEB user behavior mining of personalized recommendation

Xufang Li[1, 2*]

[1] *School of Economics and Management, Tongji University, Shanghai, 200092, China*

[2] *School of Management, Shanghai University of Engineering Science, Shanghai, 201620, China*

Received 6 October 2014, www.cmnt.lv

Abstract

Personalized recommendation directly decides the result set pushed to users and affects the quality of personalized information service. And analysis of user behavior is the key to realize personalized recommendation. The paper studies the user behavior mining based on content and VecPat-tree. When mining based on content, compound word judgment is joined in segmentation process, and the concept of keyword position factor is added to keyword weight calculation. When mining based on VecPat-tree, the paper proposed the algorithm based on VecPat-tree to process user behavior mining. The algorithm based on VecPat-tree uses the strategy of binary tree growth to avoid unnecessary projected database and effectively distinguish distribution and partial support. The paper simulated 193000 browse records of users in the experimental database to compare PrefixSpan algorithm and the algorithm based on VecPat-tree in many aspects, such as running time. And the experimental results show that the algorithm based on VecPat-tree can be more effective than PrefixSpan algorithm in achieving personalized recommendation to improve the quality of personalized information service.

Keywords: User Behavior; Data Mining; Sequence Pattern; Personalized Recommendation

1 Introduction

As a kind of service concept, personalized information service is put forward in the field of information service is not accidental. It is developed with the computer technology, network technology and the modern information processing technology. In recent years, personalized information service has always been a research hot spot.

A large number of theoretical studies and practical application proved that the implementation of personalized recommendation and optimization of personalized service strategy have a positive effect to improve customer satisfaction and loyalty to the enterprise, and to increase profit and core competitiveness of the enterprise [1-3].

2 User behaviour mining based on content

2.1 CALCULATION OF THE WEIGHTED WORD

Keyword weight calculation is not merely a word frequency statistics, but also weights their importance according to their different positions. This method is called weighted word frequency statistics.

Chinese page format on internet is HTML mostly. According to HTML tags, keywords' position in the page can be judged, thus importance of keywords in the page can be judged too.

Tag value is represented by P_k, and tag type is represented by k. Assumed: P_0 =10 (Title), P_1=5 (Primary Title), P_2=4 (Secondary Title), P_3=3 (Third Title), P_4=1(Body).

So W_c, that is weighted word frequency of each keyword T in the page, can be calculated by formula (1):

$$W_c = \sum_{k=0}^{4} Num_k \times P_k \quad , \tag{1}$$

where W_c is the weighted word frequency, Num_k is the number of keyword T's occurrence in the first k type tag; P_k is the weight of the first k type tag.

Keywords are derived from page feature vectors. Keyword weight W should be obtained by recalculating combined W_c and W_a .Weight W of keyword T in each page feature vector can be calculated by formula (2):

$$W = W_c + W_a \quad , \tag{2}$$

where W is the Weight of keyword T, W_c is the weighted word frequency obtained by content mining, W_a is the page behavior weight obtained by behavior mining.

After calculating the keyword weight, it is need to integrate all pages feature vectors to form the final user behaviour model. User's interest for a period of time basically is fixed, therefore in the integration of all page keywords, keywords will inevitably occur repeatedly. Thus for any two page feature vectors, if they have the

[*] ***Corresponding author*** e-mail: lucylxf@163.com

same keywords, the weights of the keywords should be accumulated.

2.2 CALCULATION OF COMPOUND WORDS

When dividing the words of the document, a compound word is cut into two words frequently. For example, "data mining" will often be cut into "data" and "mining" two words. Therefore, it is necessary to use a compound word generation algorithm after dividing words.

Using formula (3) to calculate the probability of two words composing a compound word:

$$f = f_{xy} \Big/ \ f_x + f_y - f_{xy} \ , \tag{3}$$

where f_x is the word frequency of word X, f_y is the word frequency of word Y, f_{xy} is the word frequency of word XY.

When f> 1%, XY should be a compound word. Thus the segmentation result will be modified and XY is recorded as a phrase.

3 User behavior mining based on VecPat-tree

Sequence pattern mining problem is put forward by Agrawal and Srikant for the first time in 1995 [14]. Many application areas are related to the sequence pattern mining [15]. While browsing a Web page, the user behavior pattern can be seen as a sequence pattern. By mining user behavior, the user's browsing law can be drawn, and his interests and hobbies can be explored.

3.1 BASIC CONCEPT

User behavior database D is collection of <ID, S> tuples, as table 1. In table 1, ID is user ID, and S is user behavior pattern.

If α is a subsequence of S, the tuple <ID,S> includes α. The support of α in user behavior database is the number of α, referred to as supports(α). min_sup is a positive integer, and it represents the minimum support threshold. If supports(α) ≥ min_sup, α is considered to be frequent in database D.

TABLE 1 User behavior database D

ID	d_1	d_2	d_3	d_4	d_5	d_6
1	1	1	0	1	0	1
2	1	0	1	0	1	1
3	1	0	0	1	1	1
4	1	0	0	0	1	1

Definition 1 VecPat-tree A vector pattern tree has to be defined as a rooted tree T=<V(T),E(T)>. Wherein, V(T) is the set of nodes, each node v represents a (or W) value and support distribution of the sequence element, denoted as v.value and v.sup. E(T) is a (directed) edge set, $E(T) \subset V(T) \times V(T)$. A constraint condition: $\forall v \in V(T)$, $v.\sup \geq \min\_\sup$. A VecPat-tree is built according to user behavior database. The path of from its root node to each leaf node is a sequence pattern.

3.2 USER BEHAVIOR MINING ALGORITHM BASED ON VECPAT-TREE

According to the above concepts, user behavior mining algorithm based on VecPat-tree is given below.

Algorithm: use VecPat-tree to mine user behavior sequence patterns through pattern growth.

Input: user behavior database D; min_sup: the minimum support threshold.

Output: user behavior pattern set.

Step 1: To reduce the number of I/O, input user behavior sequences in database D to the hash table V.

Step 2: Build a VecPat-tree.

1) At first build the root node of the VecPat-tree, marked with 1.

2) Respectively build left child node and right child node of the node, marked with 0 and 1. Push the path from the node to the root node into a stack. Then pop the stack and get a sequence s_i. s_i-sup(the node support) = s_i-sup+1 if there is s_i sequence in the hash table V. If s_i-sup<min_sup in a subtree, delete the subtree and mark the left or right child node of the node with NULL. If left child node and right child node of the node are all NULL, the node is a leaf node. Then put the node into a queue which used to store all leaf nodes.

3) Recursively implement

4) until all leaf nodes in the VecPat-tree are put into the queue.

Step3: Generating user behavior pattern set.

```
Stack s ;  //initial stack s
for each leaf node si(si ∈ Queue) do
   repeat
      s.add(si)  // push stack
      si= si+1   // back to the parent node
   until si == null
   repeat
      sF += s.pop( ) // generate sequences
   until s is empty
end for
Result = UsF
```

Step 4: Remove the end containing 0 of the sequence, and then judge whether the deleted sequence is prefix of other sequences in sequence patterns set. If so, the sequence is deleted to remove redundancy. At last, output user behavior patterns set.

4 Experiment

The paper simulated 193000 browse records of users who visited a WEB page more than 20 times within 100 days. The records are put in the experimental database.

PrefixSpan algorithm and the algorithm based on VecPat-tree are compared here. The experimental results show that number and length of sequences in sequence pattern result sets obtained by the algorithm based on VecPat-tree are significantly higher than PrefixSpan algorithm, especially in the case where the threshold value is smaller. And sequence pattern result set of the algorithm based on VecPat-tree contains sequence pattern result set of PrefixSpan algorithm. Because PrefixSpan algorithm missed repeat sequence patterns containing longer interval generated by user periodic visit. However, the algorithm based on VecPat-TREE can well dig out such repeat sequence patterns. All of above are shown in Figure 1, Figure 2, and Table 2.

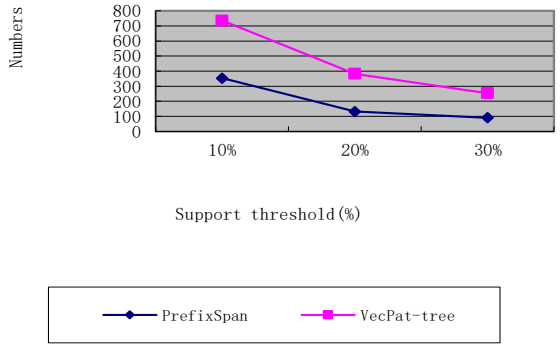

FIGURE 1 Comparison of sequence pattern number under different support

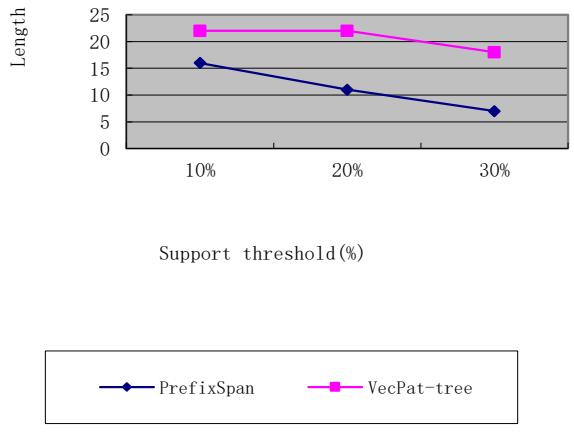

FIGURE 2 Comparison of sequence pattern length under different support

TABLE 2 Comparison of sequence pattern number and length under different support

Support	Algorithm based on VecPat-tree	PrefixSpan
10%	Length: 22 Numbers: 736	Length: 16 Numbers: 353
20%	Length: 22 Numbers: 382	Length: 11 Numbers: 131
30%	Length: 18 Numbers: 253	Length: 7 Numbers: 90

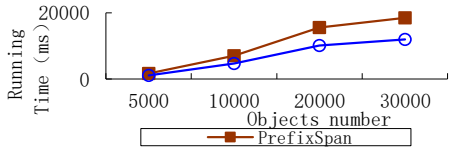

FIGURE 3 Comparison of running time at the same support (2)

Two algorithms are applicable in the sparse and dense data sets, and their advantages are more obvious in the dense data set. They can effectively to compress the data processing, and do not produce the candidate set. When need to add constraint conditions, two algorithms both can be extended easily. However running time of PrefixSpan algorithm is more than the algorithm based on VecPat-tree at the same support as shown in figure 3.

5 Conclusions

How to help people obtain the information needed from the overload information has become a problem to be solved. A personalized information service provides a new way to solve this problem. Personalized recommendation directly affects the quality of personalized information service.

The paper studies the user behavior mining based on content and VecPat-tree. When mining based on content, compound word judgment is joined in segmentation process, and the concept of keyword position factor is added to keyword weight calculation. When mining based on VecPat-tree, the paper proposed the user behavior mining algorithm based on VecPat-tree, which uses the strategy of binary tree growth to avoid unnecessary projected database and effectively distinguish distribution and partial support. Also, the experimental results show that number and length of sequences in sequential pattern result sets obtained by algorithm based on VecPat-tree are signifycantly higher than PrefixSpan algorithm, especially in the case where the threshold value is smaller. At the same support, running time of PrefixSpan algorithm is more than the algorithm based on VecPat-tree. In addition, it is more difficult to realize PrefixSpan algorithm than the algorithm based on VecPat-tree.

Therefore, the algorithm based on VecPat-tree can be more effective in achieving personalized recommendations to improve the quality of personalized information services.

Acknowledgement

The research work was financially supported byHumanities and Social Science Fund of Ministry of Education (13YJCZH122) and Course Construction of SUES (Z20140303, P201403002).

References

[1] Shankar V, Smith A K, Rangaswamy A 2003 Customer satisfaction and loyalty in online and offline environments, *International Journal of Research in Marketing* **20**(2) 153-75

[2] Souitaris V, Balabanis G 2007 Tailoring Online Retail Strategies to Increase Customer Satisfaction and Loyalty *Long Range Planning* **40**(2) 244-61

[3] Roh T H, Ahn C K, Han I 2005 The priority factor model for customer relationship management system success *Expert Systems with Applications* **28**(4) 641-54

[4] Chen Z, Wang L 2010 Personalized product configuration rules with dual formulations: A meihod to proactively leverage mass confusion, *Expert Systems with Applications* **37**(1), 383-392

[5] Changchien S W, Lee C-F, Hsu Y-J 2004 On-line personalized sales promotion in electronic commerce *Expert Systems with Applications* **27**(1) 35-52

[6] Schneider S, Shabalin P, Bichler M 2010 On the robustness of non-linear personalized price combinatorial auctions,*European Journal of Operational Research* **206**(1) 248-59

[7] Poulin M, Montreuil B, Martel A 2006 Implications of personalization offers on demand and supply network design: A case from the golf club industry *European Journal of Operational Research* **169**(3) 996-1009

[8] Cheung K-W, Kwok J T, Law M H, Tsui K-C 2003 Mining customer product ratings for personalized marketing *Decision Support Systems* **35**(2) 231-43

[9] Amstel van P, Eijk van der P, Haasdijk E, Kuilman D 2000 An interchange format for cross-media personalized publishing, *Computer Networks* **33**(1-6) 179-95

[10] Weng S-S, Liu M-J 2004 Feature-based recommendations for one-to-one marketing *Expert Systems with Applications* **26**(4) 493-508

[11] Acquisti A, Varian H R 2005 Conditioning prices on purchase history *Marketing Science* **24**(3) 367-81

[12] Lang K R, Shang R D, Vragov R 2009 Designing markets for co-production of digital culture goods, Decision Support Systems **48**(1) 33-45

[13] Montgomery A L, Smith M D 2009 Prospects for Personalization on the Internet *Journal of Interactive Marketing* **23**(2) 130-7

[14] Agrawal R, Srikant R 1996 Mining sequential patterns: Generalizations and performance improvements *EDBT*

[15] Pei Jian, Han Jiawei, Mortazavi-Asl B, Pinto H, Chen Qiming, Dayal U, Hsu Mei-Chun Pei 2001 PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth *Proceedings of Text Mine - First SIAM International Conference on Data Engineering*

[16] Xiong Y, Zhu Y Y 2007 BioPM: An efficient algorithm for protein motif mining *Proceedings of* ICBBE*'07* IEEE Press

Author

Xufang Li , 1980.06, Changsha City, Hunan Province, P.R. China

Current position, grades: The Doctoral Candidate of School of Economics and Management, TongJi University, China; the Associate Professor of School of Management, Shanghai University of Engineering Science, China.

University studies: B.Sc. in Computer Science from Nanchang University in China and received M.Sc. in Computer Science from South China Normal University in China.

Scientific interest: Her research interest fields include Information Management and Computer Science

Publications: more than 10 papers published in various journals.

Experience: She has teaching experience of 10 years, has completed 3 scientific research projects.