A HowNet Based Web Log Mining Algorithm

Chen Li, Jiayin Qi and Huaying Shu

School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, P.R. China
china.lichen@gmail.com ssfqjy@263.net shuhy@bupt.edu.cn

Abstract. Web log mining is used to extract user access pattern. An algorithm is proposed in this paper to resolve the problem of bad explanation for page sequence of web log mining. The algorithm firstly transforms user visited page sequence into maximal forward sequence, and then uses HowNet based semantic similar algorithm to describe user interest in visit sequence and explains the interest movement with certain semantic words. The algorithm will help web sites provide personalization pages.

Keywords: *HowNet, Web log mining, User navigation patterns*

1. INTRODUCTION

Agrawal et al. [1] uses sequence pattern data mining technique to extract web navigation pattern from web logs, this technique is based on the mining of association rules. Chen et al. [2] introduce the concept of Maximal Forward References which tries to divide the user's task into several transactions to extract user access pattern. This technique firstly converts the original sequence of web log data into a set of maximal forward references. By doing so, it filters out the effect of some backward references, which are mainly made for ease of traveling and concentrate on mining meaningful user access sequences. Secondly, this technique derives algorithms to determine large reference sequences from the maximal forward references obtained. The second step is based on some hashing and pruning techniques. MA Xi-jun et al. [3] considers the web users as artificial ants, and uses the ant colony approach for user navigation patterns. HE Bo et al. [4] proposes an intelligent information recommendation algorithm based on user model clustering (IRUMC). IRUMC clusters similar user models, produces user clustering models and gains recommendation sets.

These methods have the same disadvantage: they consider the high-frequency accessed path and access pattern as the user interest path. However, different pages on the path reflect the changing process of user interests. The results of these methods are just paths and access patterns which could not explain clearly how user interests are changing. They also provide very little direct suggestions for improving the web sites.

Based on these researches, this paper proposes an algorithm which firstly extracts user's access sequences, secondly analyzes the relationships between pages in these sequences, thirdly compares the key words' similarities based on the *HowNet* word

Please use the following format when citing this chapter:

Li, C., Qi, J., Shu, H., 2007, in IFIP International Federation for Information Processing, Volume 255, Research and Practical Issues of Enterprise Information Systems II Volume 2, eds. L. Xu, Tjoa A., Chaudhry S. (Boston: Springer), pp. 923-931.

semantic similarity computing algorithm, fourthly describes the changing process of user interests with key words, finally guides web sites offer customized pages for different users.

2. HOWNET

HowNet is a network system of knowledge [5]. It describes the concepts inside Chinese and English words and describes the relationships between concepts and the relationships between concepts' attributes. In HowNet, *semantic elements* are the basic unseparated elements which are extracted from all Chinese words to describe other words. Every concept is explained and expressed with a set of semantic elements.

HowNet describes a concept with a record, for instance:

NO.=017144
W_C=Da (a Chinese word)
G_C=V
E_C=~volleyball, ~poker, ~swing, ~Tai Chi, ~well in football
W_E=play
G_E=V
E_E=
DEF=exercise|Duan Lian (a Chinese word with same meaning as exercise) sport|Ti Yu(a Chinese word with the same meaning as sport)

In *DEF* partition, HowNet uses these symbols, " , ~ ^ # % $ * + & @ ? ! {} () [] " , to describe the interconnection between semantic elements. There are substantives and empty words in HowNet, this paper only discusses substantives. According the classification in 6, there are three types semantic description for substantives: (1) independent semantic element description, such as "DEF=a basic semantic element" and "DEF= (a Chinese word)"; (2) relationship semantic element description, such as "relationship semantic element=basic semantic element" or "relationship semantic element=(a Chinese word)" or "(relationship semantic element=(a Chinese word)"; (3) symbol semantic element description, such as "a symbol and a basic semantic element" or "a symbol(a Chinese word)".

There are 8 relationships between semantic elements: up-down, synonymy, antonym, sibling, attribute-owner, part-whole, material-product and event-role. These relationships combine a complex network structure. Up-down is the most important relationship. With up-down, all semantic elements could be turned into a tree structure which is the foundation for word semantic similarity computing.

3. RESEARCH FRAMEWORK

3.1 Search for Maximal Forward References

Each page visit can be described as a source-destination pairs: (s, d). s is the URL on quoted page, and d is the URL on requested page. For the references in the first requested pages, its s domain is null (empty). According to the user ID through reordering the Web log, it receives all user references $(\{(s_1,d_1),(s_2,d_2),...,(s_n,d_n)\})$, (s_i,d_i) is sorted by time. Based on the MF algorithm concept raised by Chen et al. [2], the various users' references seed the maximal forward references. Supposed that the J user's maximal forward references include k sub-references, the maximal forward reference is

$$MF_j = \{(p_{11}p_{12}...p_{1m}),(p_{21}p_{22}...p_{2m}),...,(p_{k1}p_{k2}...p_{km})\}.$$

3.2 Build Association Matrix of Pages

Definition 1 assuming a website with n pages, *Associatoin Matrix* is the association vector from quoted pages to requested page, all association vectors constitute a correlation matrix.

	p_1	p_2	...	p_n
p_1	v_{11}	v_{12}	...	v_{1n}
p_2	v_{21}	v_{22}	...	v_{2n}
...	...	...	...	...
p_n	v_{n1}	v_{n2}	...	v_{nn}

Figure 1. Association Matrix

To describe the relationship between the pages, we use the TF-IDF technology. It is the most mature and the most successful text learning technology in the field of information retrieval. The basic idea is to take a document as a vector in vectors space. Each dimension of the vector is composed by a word and its weight. Based on the document vector using TF-IDF technology to analyze the document, page title dimension and link dimension are added, because they are major reasons which affect customers switch pages.

Definition 2 Association vector v_{ij} composed by the page title t_j the link word a_{ij} from quoted page p_i to requested page p_j and the homepage keyword vector k_j. They are obtained respectively from <title> labels in the HTML document, <a> labels and the analysis result of TF-IDF technology. In particular, if $i = j$, then $v_{ij} = \{t_j, k_j\}$.

$k_j = \{word_1, word_2, ..., word_n\}$ of v_{ij} is the first n words, according to the weight $weight_i$ of every word $word_i$ ranking from high to low, if the word is same

as the page title t_j or the link words a_{ij}, it will be removed from k_j. The method of weight calculation is $weight_i = TF(word_i, p_j) \times \log \frac{|D|}{DF(word_i)}$. $TF(word_i, p_j)$ is the term frequency that the $word_i$ appears in the pages; $\log \frac{|D|}{DF(word_i)}$ is the inverse document frequency of the $word_i$, it means the more number of the pages including $word_i$, the less effect that $word_i$ distinguishes pages; $|D|$ is the total number of pages; $DF(word_i)$ is the frequency of pages, which is the number of document that $word_i$ appears at least one time in it.

3.3 Word Semantic Similarity Algorithm Based on HowNet

Liu Qun et al. propose a semantic similarity computing algorithm based on HowNet to compare the similarity between words 4. This paper improves this algorithm.

Definition 3 Assume two words $word_1$ with n concepts: S_{11}, S_{12}, ..., S_{1n}, and $word_2$ with m concepts: S_{21}, S_{22}, ..., S_{2m}, the similarity between $word_1$ and $word_2$ is the maximal similarity between their concepts:

$$Sim_1(word_1, word_2) = \max_{i=1..n, j=1...m} Sim_2(S_{1i}, S_{2j})$$

Definition 4 Assume two semantic elements S_1 and S_2, their similarity is determined by their semantic distance d in HowNet and adjusted parameter α:

$$Sim_2(S_1, S_2) = \frac{\alpha}{d + \alpha}.$$

In 6, it is firstly to computes (1) similarity between the first independent semantic elements: $Sim_{21}(S_1, S_2)$; (2)similarity between independent semantic elements without first independent semantic elements: $Sim_{22}(S_1, S_2)$; (3) similarity between relationship semantic elements: $Sim_{23}(S_1, S_2)$; (4) similarity between symbol semantic elements $Sim_{24}(S_1, S_2)$, and secondly to multiple them together with weights to get similarity between concepts. The formula is:

$Sim_3(S_1, S_2) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} Sim_{2j}(S_1, S_2)$, and β_i is pre-determined weights, $\sum \beta_i = 1$.

This algorithm considers that the first independent semantic element has a restriction on the other semantic elements. If the similarity between first independent semantic elements is low, the other parts will have a smaller contribution to the whole similarity. But it has some disadvantages: (1) as a result of accumulated multiple, the value of other parts is too small to figure out different words together with the first

semantic element; (2) if there is no certain type of semantic element, it will provide faults. For example, compare "network" with "network" which has only the first semantic element and relationship element, the result will be lesser than 1; (3) β_i is pre-determined by experts, when new words are added into HowNet, it will be reset again and again.

This paper improves this algorithm and proposes a more practical algorithm which is also more suitable for implementation with computer. This algorithm avoids the accumulated weaken effect; solves the problem of lacking some types of semantic elements; designs an automatically generated β_i sequence to help implementation without pre-determined value of experts. This algorithm is proved a better effect after empirical research.

$$Sim_3(S_1,S_2)=\sum_{i=1}^{4}\beta_i Sim_{2i}(S_1,S_2)\ ,\quad \beta_i=\begin{cases}1-\sum_{i=2}^{n}\beta_i, & i=1\\ \dfrac{2^{n-i}}{2^n}, & 1<i<n, n\le 4\\ 0, & n\le i\le 4\end{cases}$$

, n is number of existing types.

3.4 Describe the Changing of User Interests

In the maximal forward references of user j, this paper searches for *the maximal internal similarity word of vector* (in short: MISW) and *the maximal external similarity word between vectors* (in short: MESW) to describe the changing of user interests.

Definition 5 Assume an association vector v_{ij} which has m words inside it, the internal similarity of the kth word of v_{ij} ($word_k$) is: $Sim_4(v_{ij},word_k)=\frac{1}{m-1}\sum_{n\neq k}Sim_1(word_k,word_n)$; if $word_k$'s internal similarity is: $Sim_4(v_{ij},word_k)=\max_{n=1...m}Sim_4(v_{ij},word_n)$, $word_k$ is *the maximal internal similarity word of* v_{ij} (MISW) and $Sim_4(v_{ij},word_k)$ is *the maximal internal similarity of* v_{ij} (MIS).

Definition 6 for association vector v_i with n words and v_j with m words, if the similarity between $word_{ik}$ of v_i and $word_{jk'}$ of v_j is $Sim_1(word_{ik},word_{jk'})=\max_{x=1..n,y=1...m}Sim_1(word_{ix},word_{jy})$, $word_{ik}$ is *the maximal external similarity word between* v_i *and* v_j (MESW) and

$Sim_1(word_{ik}, word_{jk'})$ is *the maximal external similarity between* v_i *and* v_j (MES).

The algorithm compares the maximal internal similarity of vectors and maximal external similarity between vectors of the users' access sequences and finds out the array of key words to describe the changing of user interests.

(1) Input user *j*'s access sequence $p_1 p_2 ... p_n$, set $i = 1$ and $arr = null$ where arr is the array of key words; β_1 and β_2 are pre-determined value.

(2) Read v_{1i} and v_{ii+1} from the interest association matrix.

(3) If the MES between v_{1i} and v_{ii+1} exceeds β_1, record the MESW $word_k$ into arr; if not, compare if MIS of v_{1i} exceeds β_2, if exceeds β_2 then record the MISW $word_{k'}$ into arr, or else record the first word $word_1$ of v_{1i} into arr.

(4) $i = i + 1$, if $i > n$ then go to step (6) else go to step (5).

(5) Read v_{i-1i} and v_{ii+1}, then go to step (3).

(6) Output arr, end.

Words inside arr describe the changing of user interest. These words form a words sequence which could semantically explain the user interests, for example, "football -> basketball -> purchase".

3.5 Provide Customized Pages

When web sites are newly built and there are few users, it is suggested to analyze the web log to extract key words of user interests, then to provide customized pages based on word semantic similarity algorithm. The algorithm is as follows:

(1) Input the array of key words for user j, arr_j, set $i = 1$, $\max = 0$, assume $p_{\max}$ as the output page, and n as the number of all this web site's pages; β_3 is a pre-determined value.

(2) Assume arr_j as an association vector, try to find the MESW between arr_j and v_{ii}. If the MES Sim_{ji} exceeds β_3 and $Sim_{ji} > \max$, set $\max = Sim_{ji}$, $p_{\max} = p_i$.

(3) $i = i + 1$, if $i > n$ then go to step (2), else go to step (4).

(4) Output $p_{\max}$.

This algorithm could dynamically build the link to $p_{\max}$ and offer user their most interesting pages. When web sites have a certain cumulated users' access data, it is

suggested to use clustering techniques to summarize the key words and then provide customized pages according to these clustering results.

4. EMPIRICAL ANALYSIS

4.1 Comparison Experiment

This paper uses 4 methods for word semantic similarity computing, the first 3 methods' data are referenced from 6: (1) only uses the first independent semantic element for computing; (2) the method which Li Sujian et al. propose [6]; (3) the method which Liu Qun et al propose [7]; (4) the method of this paper.

Table 1. Comparison Experiment

Sim	word1	Word2	semantic elements of word2	(1)	(2)	(3)	(4)
High	Man	Father	People, family, male	1	1	1	0.875
Medium	Man	Woman	People, family, female	1	0.668	0.833	0.62508
	Man	Mother	People, family, female	1	0.668	0.833	0.62508
	Man	Monk	People, religion, male	1	0.668	0.833	0.62508
	Man	manager	People, #job, officer, business	1	0.351	0.657	0.625041
Low	Man	Fish	Fish	0.347	0.009	0.208	0.2217391
	Man	Apple	Fruit	0.285	0.004	0.166	0.1821429
	Man	responsibility	responsibility	0.016	0.005	0.01	0.1342105
	Man	radiogram	machine, *communication	0.186	0.008	0.164	0.1186628
	Man	Work	Affair, $take charge	0.186	0.035	0.164	0.1186628
	Man	Happy	Attribute value, situation, good fortune, good	0.016	0.024	0.013	0.031252

There are four conclusions from comparison. Method 1 is too simple and couldn't differentiate words with high and medium similarity, for instance, "man" has the same similarity 1 with "father", "woman", "mother", "monk" and "manager". Method 2 is exact in medium similarity but poor in low similarity, for instance, "man" has the similarity 0.668 with "father", "woman", "mother" and "monk"; but "man" has mere similarity(lower than 0.01) with "fish", "apple", "responsibility" and "radiogram". Method 3 tends to consider medium similarity as high similarity, for instance, "man" has a high similarity (higher than 0.75) with "father", "woman", "mother" and "monk". Generally, method 4 is the best of all and is exact in high, medium and low similarity, for instance, if two words are not the same, the similarity is not 1, so

"man" has the similarity 0.875(lower than 1) with "father"; Also, in low similarity, "man" has the similarity 0.1821429 with "apple".

It is also concluded from the low similarity comparison that because of avoiding multiplicative and the adjustment of β_i sequence, method 4 has a higher results than method 3 and is more accurate in comparison. For instance, with method 4, the similarity between "man" and "happy"(0.1821429) is 0.0634801 higher than the similarity between "man" and "radiogram"(0.1186628) while with method 3, this number is only 0.002.

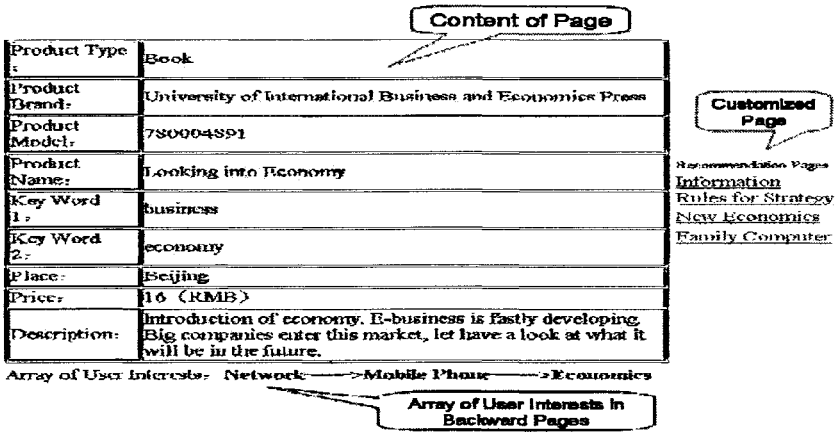

Figure1. Implementation

4.2 Implementation

This paper uses *Microsoft Visual Studio 2003* as development tool, *C#* as development language, *ACESS* as database to implement the algorithm with *B/S* framework on the platform of *Microsoft Windows Server2003* and *IIS6.0*. It could be used for recommendation system on e-business web sites. CONCLUSIONS

To solve the bad explanation problem of previous data mining methods, this paper proposes an algorithm based on HowNet word semantic similarity computing. It could describe the changing process of user interest with key words vectors. This algorithm firstly extracts users' access sequences from web log data, secondly analyzes the relationships between pages with matrix, thirdly compares the semantic similarity between key words, fourthly describes the changing process of user interests, and finally provides customized pages for every customer. After empirical analysis, this algorithm has a high accuracy and feasibility.

REFERENCES

1. R. Agrawal and R. Srikant, Mining Sequential Patterns, in *Proc. of the 11th International Conference on Data Engineering* (IEEE Computer Society Press: Washington DC, USA, 1995), pp.3-14.
2. M.S. Chen, J.S. Park, and P.S. Yu, Efficient data mining for path traversal patterns, *IEEE Trans Knowledge Data Engn.* Volume 10, Number 2, pp.209-221, (1998).
3. X. Ma, H. Ling, Y. Liu, and Y. Jiang, An Ant Colony Approach for Discovery of Users Interest Navigation Paths, *Chinese Journal of Management Science.* Volume 14, Number 3, pp.56-59, (2006).
4. B. He, W. Yang, J. Zhang, and Y. Wang, Intelligent information recommendation algorithm based on user model clustering, *Computer Engineering and Design.* Volume 27, Number 13, pp.2360-2361, (2006).
5. Z. Dong and Q. Dong, *HowNt* (March 6, 2004). http://www.keenage.com/zhiwang/c_zhiwang.html
6. S. Li, J. Zhang, X. Huang, and S. Bai, Semantic Computation in Chinese Question-Answering System, *Journal of Computer Science and Technology.* Volume 17, Number 6, pp.933-939, (2002).
7. Q. Liu and S. Li, Word Similarity Computing Based on HowNet, *Computational Linguistics and Chinese Language Processing.* Volume 7, pp.59-76, (2002).