# Personalized Recommendation of Preferred Paths Based On Web Log

Zhurong Zhou
Institute of Computer and Information Science, Southwest University, Chongqing, China
zhouzr@swu.edu.cn

Dengwu Yang
Institute of Computer and Information Science, Southwest University, Chongqing, China
abc1988@swu.edu.cn

***Abstract*—With the development of the Internet, web service generates a large amount of log information, how to mine user preferred browsing paths is an important research areas. Current researches mainly focus on the mining of user preferred browsing paths, however, they do not delve into the personalization of preferred paths and lack of semantic information. To provide personalized preferred paths to fulfill user need, in this paper, we proposed a novel method to compute the similarities of preferred paths and the given fields by experts. Firstly, the similarities of each page on the preferred paths and the given fields are computed. Secondly, according to the computed similarities of each page on the preferred paths and the given fields, the average similarity of all the pages on the preferred path and the the given field is computed, and it is used as the similarity of preferred path and the given field. After the processing mentioned above, the website can automatically recommend the related preferred paths to users according to the choice of users. Experiments show that it is accurate and scalable. It can be applied to optimize website or design personalized service.**

***Index Terms*—preferred browsing path; web usage mining; web log; matrix computing**

## I. INTRODUCTION

With the development of the Internet, web services generate a large amount of web log information everyday, how to mine the information that addresses the needs of users, provide personalized services and help the web designer update the site's topology become increasingly important.

As a part of Web mining, web log mining mainly includes sequential pattern mining[1][2][3][4] and user behavior pattern mining[5][6][7][8]. As a part of user behavior pattern mining, The mining of user preferred browsing path has become a research hotspot of web log mining. For the mining of user preferred browsing path, the commonly used algorithms include Maximal Forward References[9], reference length[10] and tree topology[11][12]. The method of Maximal Forward Reference forms several subsequence according to the characteristic of users roundtripping, the method of reference length form a number of subsequence according to the dwell time of users on web page, and the method of tree topology regard the entire log as browsing subsequence. However, these algorithms don't delve into the obtained preferred paths by now. The main limitation of above algorithms is that they do not support personalized services[13][14][15][16]. The preferred paths mined by above algorithms are lack of semantic information. When users visit a website in which a set of preferred paths has been mined, they will be confronted with a couple of path options, And no way to judge between two choices. Because the above algorithms mine preferred paths without semantical description of the attribute and the field that the path belongs to. In a practical application, each preferred path may involve a particular field. In order to provide personalized services for users, in this paper, for the obtained preferred paths, the most important t features are extracted from each page on each preferred path. After the step, according to the related fields given by expert and the field feature item set, the similarities of each page on the preferred paths and the given fields are computed. Finally, according the computed similarities of each page and the given fields, the similarities of preferred paths and given fields are computed. When users visit a website which has been mined using the method we proposed in this paper, the website can automatically recommend the related preferred paths which have high similarities with selected field according to the users's selected field.

The rest of the paper is organized as follows. Section II briefly presents the related work about the mining algorithm of the preferred paths which is proposed by Xing Dong-shan[17]. Section III detailedly introduces the method proposed in this paper to analyze the fields that the preferred paths belong to. Finally, the last two sections introduce experimental analysis and conclusion, respectively.

## II. RELATED WORK

In this paper, in order to obtain the user preferred browsing paths, we are based on the following premise: Let U denote a set of a website URL, S denote a set of all browsing sub-paths, if exists $s \subset S$ , for $\forall x \in s$ (x is a browsing page sequence composed by $u$, where $\forall u \in U$ , we call the ith browsing page as the ith bit) , their preceding m bits are the same and the (m+1)th bit has n


doi:10.4304/jsw.9.3.684-688

different choice.

### A. Preference

Suppose that the page browsing user has n different choice to leave page Q, the choice that the number of occurrences is relatively high is the user preferred pages. In the literature[17][18], the definition of support-preference can be written as following formula.

$$P_S = \frac{S_k}{\frac{1}{n}\left\{\sum_{i=1}^{n} S_i\right\}} \times S_k \tag{1}$$

Where $S_i$ denotes the support of the ith choice.

### B. Preferred Path

Web log files record the basic information of users visiting website. After preprocessing for web log files, we extract URL_R and URL, where RUL_R denotes referrer page and URL denotes navigating page. Then we establish a user access matrix using URL_R and URL, where URL_R represents the row, and URL represents the column .

#### 1) Web access matrix representation

When we establish web access matrix, we add NULL into the matrix's(Figure 1) rows and columns, In the row vector, NULL indicates that users access web page by inputting the URL in the browser or other website links, in the column vector, NULL indicates that the visit of users is end in this web page or they link to other website. In a large and actual website, the number of URLs is generally very large, if we adopt web access matrix to store the URLs, the storage space occupied by the web access matrix is proportional to the square with the number of the URLs. In fact, the URL links of each page are generally not more than a few dozen, therefore, if we adopt web access matrix, it will waste a lot of storage space. Obviously, web access matrix is a sparse matrix, hence, we adopt 3-tuples to store the non-zero elements of the web access matrix.

$$M_{(m+1)\,(m+1)} = \begin{array}{c} \\ \text{NULL} \\ \text{URL} \\ \text{URL} \\ \vdots \\ \text{URL} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{NULL} & \text{URL} & \text{URL} & & & \text{URL} \end{array} \\ \begin{bmatrix} A_{00} & A_{01} & A_{02} & \cdots & A_{0m} \\ A_{10} & A_{11} & A_{12} & \cdots & A_{1m} \\ A_{20} & A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ A_{m0} & A_{m1} & A_{m2} & \cdots & A_{mm} \end{bmatrix} \end{array}$$

Figure 1 web access matrix

#### 2) User preferred browsing path

In the representation of web access matrix, for the (i+1)th (0≤i≤m) row and the (j+1)th (0≤i≤m) column, its support-preference that is greater than or equal to the threshold value means $\left( S_{ij} / \left( \left( \sum_{k=0}^{m} S_{ik} \right) / n \right) \right) \times S_{ij} \geq P$, where (m+1) is the number of rows of matrix, n is the number of non-zero columns of the (k+1)th row and P is the threshold of support-preference.

Combine the sub-paths that support-preference is greater than or equal to the threshold value and form the user preferred browsing paths. When we combine the sub-paths, we adopt the gradual combination method, that is, the operation of combination is only carried out between the paths of the same length, every combination the length of the path increases one until the paths can not be combined. The paths which can not be combined are the user preferred browsing paths.

## III. Personalized Recommendation Based on Preferred Paths

### A. The Overall Thought

Our task is to find the field that the preferred path belongs to. In order to solve this question, we discover that , for any preferred paths, the content of the pages on the paths may be very similar or dissimilar to the given field. Thus, we proposed the following solution.

Step 1. For each obtained preferred path, we extract the most important previous t features from each page on the preferred path, and compute the similarity of each page and each field given by expert. To compute the similarity of each page and each given field, we adopt the following formula.

$$sim(Qi, Fj) = \frac{Qi \cdot Fj}{\|Qi\|\|Fj\|} \tag{2}$$

Where "$\cdot$" denotes the dot product of the vector, $\|Qi\|$ is the length of vector Qi, and $\|Fj\|$ is the length of vector Fj.

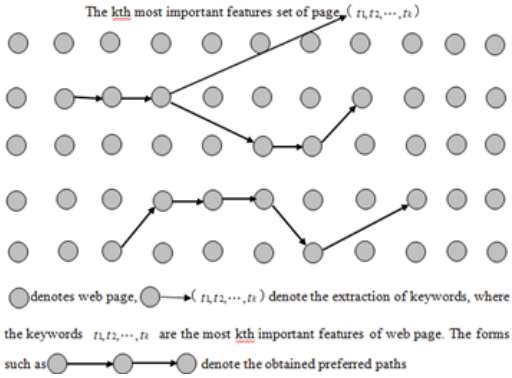


Figure 2 Paths and the most important previous k features of each page

Step 2. For each page of preferred path, according the similarity of each page and the given field, we construct a matrix shown in the table I. In the table I, F1, F2, F3, ⋯, Fm denote the related fields given by expert, and Q1, Q2, ⋯, Ql (where l is the number of pages on preferred path) denote the pages on the preferred path. For the value $a_{ij}$ of arbitrary element of the table I, it denotes the similarity of page Qi ($1 \le i \le l$) and Field Fj ($1 \le j \le m$), where $0 \le a_{ij} \le 1$.

TABLE I.

THE MATRIX REPRESENTATION OF THE SIMILARITY OF PAGES ON THE PREFERRED PATH AND GIVEN FIELDS

| | F1 | F2 | F3 | ... | Fm |
|---|---|---|---|---|---|
| Q1 | $a_{11}$ | $a_{12}$ | $a_{13}$ | ... | $a_{1m}$ |
| Q2 | $a_{21}$ | $a_{22}$ | $a_{23}$ | ... | $a_{2m}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | | ⋮ |
| Ql | $a_{l1}$ | $a_{l2}$ | $a_{l3}$ | ... | $a_{lm}$ |

Step 3. For each matrix obtained in the step 2, we compute the average similarities of all the pages and each field according to the similarities of each page Qi ($1 \le i \le l$) and each field Fj($1 \le j \le m$) and we regard the average similarities as the similarities of the preferred paths and given fields. For the given field Fj, the formula used to compute the average similarity is written as following.

$$avg = \frac{\sum_{i=1}^{l} a_{ij}}{l} \quad (3)$$

According to the average similarity, we construct a matrix shown in the table II.

TABLE II.

THE MATRIX REPRESENTATION OF SIMILARITY OF PREFERRED PATHS AND GIVEN FIELDS

| | F1 | F2 | F3 | ... | Fm |
|---|---|---|---|---|---|
| P1 | $a_{11}$ | $a_{12}$ | $a_{13}$ | ... | $a_{1m}$ |
| P2 | $a_{21}$ | $a_{22}$ | $a_{23}$ | ... | $a_{2m}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | | ⋮ |
| Pn | $a_{n1}$ | $a_{n2}$ | $a_{n3}$ | ... | $a_{nm}$ |

For the value $a_{ij}$ of arbitrary element of the table II, $a_{ij}$ denotes the similarity of the preferred path Pi ($1 \le i \le n$, where n is the number of preferred paths) and the given Field Fj ($1 \le j \le m$).

Step 4. When users visit the website, according the selected Field Fj of users, for the obtained matrix in the step 3, we sorted all the elements in the column Fj in descending. After these steps, the related preferred paths corresponding to the elements sorted ahead are recommended to users.

### B. The Description of Algorithm

For the solution proposed in the previous section, it involves two algorithms. The first algorithm is the computation of the similarity of each page on the preferred path and the given fields. The second algorithm is the computation of the similarity of the preferred paths and given field. The description of the two algorithms is following, respectively

*1) The computation of similarity of page- field*

In order to compute the similarity of each page on the preferred path and the given fields, we adopt the cosine similarity to denote the similarity of page-field. The description of the algorithm as following:

**Description of Algorithm:**

Input: M M is a $m \times r$ matrix, where m denotes the number of the given fields, and r denotes the size of field feature item set.

P P is a $l \times r$ matrix, where $l$ denotes the number of pages on preferred path, and $r$ denotes the number of features extracted from each page.

Output: R a matrix of similarity of page-field

```
1  create a l × m matrix R
2  for each row F in M
3     set j equal the number of the current row of M
4     For each row p in P
5        set i equal the number of the current row of P
6        compute the similarity sim of F and P
7        set R(i, j) =sim
8     End for
9   End for
10  return R
```

*2) The computation of similarity of preferred path and field*

In order to compute the similarity of preferred path and given field, the description of algorithm as following.

**Description of Algorithm:**

Input: S a set of matrice obtained by above algorithm.

Output: R a matrix used to store the similarities of preferred paths and given fields.

```
1  create a integer variable and set it equal the size of S
2  create a n × m matrix R
3  for each element e ∈ S // e is matrix
4     set i equal the index position of e in the set S
5     For each column of e
6        set j equal the number of the current column
7        compute the similarity sim of F and P
8        set R(i, j) =avg
9     End for
10  End for
11  return R
```

After the computation of algorithm 1) and algorithm 2), we finally obtain a $n \times m$ matrix. When users visit the website, according to selected field of users, the system automatically selects all the elements corresponding to the selected field and sort them in descending . After sorting, the system automatically recommend the preferred paths corresponding to the top elements to users.

## IV. EXPERIMENTAL ANALYSIS

In the process of our experiment, we adopt the algorithm proposed by Xing Dong-shan[17] to obtain the preferred paths. In the experiment, our experimental data comes from a web site's log data. After data preprocessing, we obtain the following web access matrix shown in table III.

According to the data shown in table III, we obtain four user browsing preferred paths, they are (NULL, A, C, NULL), (D, G, H. NULL), (NULL, A, B, F, NULL), and (NULL. A, C, G, H, NULL). However, in the experiment, The NULL on each preferred path has no effect for the follow-up experiment, so we delete all NULL on the whole preferred paths and the final user browsing preferred paths are denoted as (A, C), (D, G, H) (A, B, F) and (A, C, G, H).

For the pages of arbitrary preferred path, we adopt the method ID_TDF to extract the keywords from each page. For arbitrary preferred path, we compute the similarity of each page and each given field according to the given fields and related field feature item set. In the process of the experiment, according to the given fields and related

field feature item set by expert, for the preferred paths (A, C), (D, G, H), (A, B, F), and (A, C, G, H), the similarities of pages and given fields are shown from table IV to table VII, respectively. According to the computed similarities, the similarities of preferred paths and given fields are computed by using the algorithm 2) and shown in table VIII.

TABLE III.

WEB ACCESS MATRIX

| | NULL | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|---|
| NULL | 0 | 35 | 5 | 1 | 2 | 0 | 0 | 0 | 1 |
| A | 5 | 0 | 20 | 10 | 0 | 0 | 0 | 0 | 0 |
| B | 3 | 0 | 0 | 0 | 5 | 0 | 12 | 5 | 0 |
| C | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 1 |
| D | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 5 | 0 |
| E | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 14 |
| H | 10 | 0 | 6 | 6 | 0 | 0 | 0 | 0 | 0 |

TABLE IV.

THE SIMILARITIES OF ALL PAGES OF PREFERRED PATH (A, C) AND GIVEN RELATED FIELDS

| | F1 | F2 | F3 |
|---|---|---|---|
| A | 0.8 | 0.1 | 0.1 |
| C | 0.6 | 0.1 | 0.3 |

TABLE V.

THE SIMILARITIES OF ALL PAGES OF PREFERRED PATH (D, G, H) AND GIVEN RELATED FIELDS

| | F1 | F2 | F3 |
|---|---|---|---|
| D | 0.6 | 0.3 | 0.2 |
| G | 0.7 | 0.2 | 0.1 |
| H | 0.3 | 0.4 | 0.3 |

TABLE VI.

THE SIMILARITIES OF ALL PAGES OF PREFERRED PATH (A, B, F) AND GIVEN RELATED FIELDS

| | F1 | F2 | F3 |
|---|---|---|---|
| A | 0.2 | 0.6 | 0.1 |
| B | 0.2 | 0.5 | 0.2 |
| F | 0.4 | 0.6 | 0.3 |

TABLE VII.

THE SIMILARITIES OF ALL PAGES OF PREFERRED PATH (A, C, G, H)AND GIVEN RELATED FIELDS

| | F1 | F2 | F3 |
|---|---|---|---|
| A | 0.7 | 0.1 | 0.2 |
| C | 0.6 | 0.1 | 0.3 |
| G | 0.3 | 0.4 | 0.3 |
| H | 0.6 | 0.3 | 0.1 |

TABLE VIII.

THE SIMILARITIES OF PREFERRED PATHS AND GIVEN FIELDS

| | F1 | F2 | F3 |
|---|---|---|---|
| (A, C) | 0.7 | 0.1 | 0.2 |
| (D, G, H) | 0.5333 | 0.3 | 0.2 |
| (A, B, F) | 0.2667 | 0.5667 | 0.2 |
| (A, C, G, H) | 0.55 | 0.225 | 0.225 |

From table III to table VII, we can see that, for the pages on the preferred path, the degrees of the different fields that the pages belong to are often different in general. In a real word, each page may include different fields content, and the degrees of different fields that the page belongs to are often different. Thus, the method proposed in this paper can refect this situation well. By the experiment, we can see that, when users visit the website mined by the method proposed in this paper, the system can automatically recommend the related preferred paths to the users according to the selected field of users.

For the algorithm proposed by Xing Dong-shan, it can effectively mine user browsing preferred paths, but the preferred paths themselves are lack of sematic information, they are very hard to fulfill web users' need. However, the method proposed in this paper makes up the defects. When users visit a web site mined by the method proposed in this paper, the system can automatically recommend related preferred paths for users according to users' choice.

## V. CONCLUSION

Currently, in the web usage mining, most methods or algorithms mainly focus on the mining of the user preferred browsing paths. For the obtained user preferred browsing paths, these researches don't make a deep analysis. Because of the reason, it is very hard to provide personalized services for users. In this paper, we proposed a novel method which can be used to analysis the fields that the obtained preferred paths belong to. The method proposed in this paper can compute the similarities of preferred paths and given different fields by expert. It means the degrees of different fields that the preferred paths belong to are often different. To achieve personalized services, when users access the website mined by the method proposed in this paper, the system can automatically recommend the preferred paths which have high similarities with the selected field according to the selected field of the users.

In this paper, for the obtained preferred paths, we only analyze the fields that the preferred paths belong to. In the future work, we can also use other tools or method[, such as literatures[19][20][21] showing to research preferred paths.

## ACKNOWLEDGMENT

My mentor gives me much help during the process of completing this paper. In the whole process, he gave me a lot of advice, without his help, it should be more difficult

for me to complete this paper. I take the opportunity to my mentor to express my heartfelt thanks and highest respect.

## REFERENCES

[1] Hart Jiawei, Pei Jian, Yan xifeng. From Sequential Patten Mining to Structured Pattern Mining: A Patten — Growth Approach [JJ. Journal of Computer Science and Technology, 2004, 19:257—279.

[2] Agrawal R, Srikant R. Mining sequential patten [A]. Proc 1995 Int Conf Data Engineering (ICDE's95) [C].Chinese Taipei: [s.n.],1995.3—14.

[3] Srikant R, Agrawal R. Mining sequential patten; Generalizations and performance improvements [A]. Proc 5th Int Cord Extending Database Technology (EDBT's96) [C].Avignon. France:[S.n.],1996.3—17.

[4] Pei Jian, Han Jia—Wei, Mortazavi—Asl B, et al. Mining sequential pattern by Pattern —Growth: The Prefix Span Approach [J].IEEE TKDE, 2004, 16(10):9—13.

[5] Perkowitz M., Etzioni O.. Towards adaptive sites: Conceptual framework and case study. Artificial Intelligence,2000,118:245～275

[6] Schechter S., Krishnan M., Smith M.D. Using path profiles to predict HTTP requests. In: Proceedings of the 7th International World Wide Web Conference Computer, Networks and ISDN System, Brisbane, Australia, 1998, 30: 457～467

[7] Cao, L.: In-depth Behavior Understanding and Use: the Behavior Informatics Approach Information Science 180(17), 3067–3085 (2010)

[8] V.V.R. Maheswara Rao, Dr. V. Valli Kumari, Dr. KVSVN Raju. Understanding User Behavior using Web Usage Mining. International Journal of Computer Applications. 2010.

[9] Chen M S, Park J S. Data mining for path traversal patterns in a Web environment. In: Proceeding of the 16th International Conference on Distributed Computing System, Hong Kong,1996.385-392

[10] Chen M S, Park J S. Data mining for path traversal patterns in a Web environment. In: Proceeding of the 16th International Conference on Distributed Computing System, Hong Kong,1996.385-392

[11] Ratnesh Kumar Jain, Dr.R.S. Kasana. Efficient Web Log Mining using Doubly Linked Tree. International Journal of Computer Science and information Security, 2009.

[12] Ahmed, C.F., Tanbeer, S.K., Jeong, B.-S., Lee, Y.-K.: Efficient Tree Structures for High Utility Pattern Mining in Incremental Databases. IEEE Transaction on Knowledge and Data Engineering 21(12), 1708‑1721 (2009)

[13] Zui Zhang, Hua Lin, Kun Liu, Dianhuang Wu, Guangquan Zhang, Jie Lu. A hybrid fuzzy-based personalized recommender system for telecom products/services. Information Sciences, 235, p.117-129, Jun 2013

[14] M.A. Ghazanfar, A. Prügel-Bennett, S. Szedmak. Kernel-mapping recommender system algorithms. Information Sciences, 208(2012), pp.80-104

[15] W. IJntema, F. Goossen, F. Frasincar et al., Ontology-based news recommendation, In proceedings of the 2010 EDBT/ICDT Workshops, Lausanne, Switzerland,, 2010, pp.1-6

[16] J. Lu, Q. Shambour, Y. Xu *et al*. BizSeeker: a hybrid semantic recommendation system for personalized government-to-business e-services. Internet Research, 20(3) (2010), pp.342-36

[17] Xing Dong-Shan, Shen Jun-Yi, Song Qin-Bao. Discovered Preferred Browsing Path from Web Logs. Chinese Journal of Computer, Vol. 26, No. 11, Nov. 2003

[18] He Yue, Chen Dayong, Teng Ge'er. Research user preferred browsing paths based on web data mining. Computer Engineering and Application, 2012, 48(7):106-108

[19] Guojun Ding, Lide Wang, Peng Yang, Ping Shen, Shuping Dang. Diagnosis Model Based on Least Squares Support Vector Machine Optimized by Multi-swarm Cooperative Chaos Particle Swarm Optimization and Its Application. Journal of Computers, Vol 8, No 4 (2013), 975-982, Apr 2013

[20] Ling Chen, Yixin Chen, Li Tu. A Fast and Efficient Algorithm for Finding Frequent Items over Data Stream. Journal of Computers, Vol 7, No 7 (2012), 1545-1554, Jul 2012

[21] Xin Wan, Qimanguli Jamaliding, Toshio Okamoto. Analyzing Learners' Relationship to Improve the Quality of Recommender System for Group Learning Support. Journal of Computers, Vol 6, No 2 (2011), 254-262, Feb 2011

**Zhurong Zhou**, he was bord in 1970. Now he is an associate professor in Southwest University. His main research interests include data mining, semantic network, s search engine and online education.

**Dengwu Yang** was born in Bijie City, Guizhou Province, China, in 1988. Now he is a graduate student in Southwest University. His research interests include Web data mining and pattern recognition.