ISSN: 0976-3031

Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 8, Issue, 7, pp. 18712-18716, July, 2017

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

A METHOD BASED ON SEQUENTIAL ACCESS PATTERNS FOR INFORMATION RETRIEVAL

Poonam Yadav*

D.A.V College of Engineering & Technology, Kanina, Haryana 123027, India

DOI: http://dx.doi.org/10.24327/ijrsr.2017.0807.0554

ARTICLE INFO

Article History:

Received 16th April, 2017
Received in revised form 25th May, 2017
Accepted 23rd June, 2017
Published online 28th July, 2017

Key Words:

Information Retrieval, web usage mining, sequential access patterns, Pattern-tree

ABSTRACT

Extraction of relevant information from the web has become the emerging challenge as the data in World Wide Web has been progressively increased. To solve this problem, a valid method called web usage mining is used that can mine weblogs for user models and recommendations. The general recommender systems of the web mainly function on the basis of association rule mining and clustering. Apart from this, web personalization is proposed in this paper which adopts sequential access pattern mining. The recurrent sequential web access patterns are recognized through this method. Those patterns are further stored in tree structure termed as pattern-tree. The subsequent processes such as matching and producing web links for recommendation are done with these stored patterns. Finally, a valuable performance analysis is carried out to validate the proposed model.

INTRODUCTION

Since the data accessible on World Wide Web has been growing periodically, it has been much more complex to retrieve the required information from the web [24] [25] [26] [27] [28]. Under such circumstance, a technique termed as web personalization has developed to make easy to extract the relevant information [9] [3] [1] [10] [11] [12]. Moreover, the current study concerns on establishing an intelligent recommender system for the purpose of accessing related web pages effectively through personalized web service. The task in determining which web pages are more probable to be accessed by the operator in the future is the main objective of intelligent recommender method [13] [14] [15] [16] [17].

Subsequently, two methods have been developed for guiding the web recommendations are hybrid content-based collaborative filtering [7] and collaborative filtering [8] techniques. As more users are browsing the websites simultaneously in a secret way, this method has failed to attain their reliability. Thus it is difficult to obtain their concealed identities [18] [19] [20] [21] [22] [23]. Recently, a number of methods based on web usage mining have developed to determine to simulate usage forms obtained from the information deposited in web browser web server logs. Furthermore, the web recommendation has been applied with auspicious web usage mining approaches like clustering [5] and association rule mining [6].

This paper proposes an intelligent web recommender system, which adopts sequential pattern mining approach. In fact, the proposed method is completely dissimilar from almost all of the traditional web recommendation methods. The sequential pattern mining methods consider the sequential characteristics of access patterns [2], which cannot be found in association rule mining and clustering. Moreover, the proposed method can foresee the subsequent web pages. In addition, a compact data model termed as Pattern-tree is also developed by this paper that stores the sequential web access patterns. It is a competent method for matching the user patterns and generation of recommendation rules. Ultimately, the performance the recommender system is measured through analyzing diverse evaluation measures such as satisfaction, precision, and availability.

This paper contributes to developing an effective information retrieval system using Sequential access pattern. The rest of the paper is organized as follows. Section II provides the architecture of the proposed system. Section III illustrates the overall methodology. Section IV portrays the performance evaluation. Section V concludes the paper.

Architecture of System

The proposed system architecture for information retrieval is shown in Fig. 1. Initially, the WWW server of the website records the web access activities of the website. Further, it is stored in the Web Server Logs. The details such as IP address

Corresponding author:* **Poonam Yadav
D.A.V College of Engineering & Technology, Kanina, Haryana 123027, India

of the client, requested URL, access time, ID of the user, status code of HTTP, etc. are recorded by each user access. Subsequently, the user access sequences from the Web Server Logs are mined by the relevant components of sequential pattern mining. With the mined sequential web access patterns, the Pattern-tree is constructed by the patter-tree construction component. Moreover, the processes associated with the sequential pattern mining and pattern-tree construction are carried out offline. The new access data can be integrated by updating the patter-tree periodically.

During the website visits, the HTTP of the user request in the present browsing session are saved in a well-organized manner and further build up the present access sequence. The IP address of the user is essential to recognize the website accessing by the user. Furthermore, the recommendation rules can be generated by the generation component by equalizing the present access sequence of the user from the recommendation model of the pattern-tree. Then the related links are animatedly injected into the present requested page from the extracted recommendation rules.

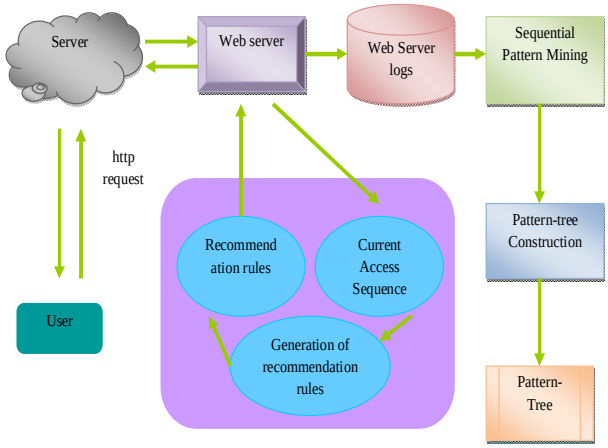

Fig 1 Architecture of proposed Information Retrieval System

METHODOLOGY

Sequential Pattern Mining

The fundamental information such as IP address of the client, requested URL, accessed time, ID of the user, status code of HTTP, etc. are available in each line of weblogs. In sequential pattern mining, a group of sequences is determined in the case of defining the weblogs. These sequences consist of web access events from each user at the time of associated session in timestamp ascending order. Before starting the sequential pattern mining, it is needed to carry out the pre-processing on the weblogs. Those pre-processing techniques include user identification, data cleaning, and session identification.

This paper describes the sequential pattern mining as depicted as follows. Assume a sequence database, where each sequence is a group of transactions. These transactions are well arranged by transaction time. Here, each transaction consists of a group of items, which can recognize the entire sequential patterns user-specified minimum support. A number of data sequences that hold the pattern are the definite manner of the user-specified minimum support.

Consider a group of exclusive access events that is denoted as D. The access events access the web resources such as URLs, web pages, themes. Let $X = z_1 z_2 .... z_n$ be the access sequence, where $z_i \in D$ for $i = 1,2....n$. The respective sequence is an ordered assortment of access events. Moreover, the length of the web sequence is denoted as l, determined by $l = |X|$. Note that it is not essential that $z_j \neq z_i$ for $i \neq j$ in X. Also, consider a group of events $D = \{i, j, k, l, m, n\}$ is present in web access sequence. Accordingly, Table 1 shows the sample web access sequence databases that contain four web access sequences.

A web access sequence $X' = z_1 ' z_2 ' .... z_n$ is termed as web access sequence $X = e_1 e_2 .... e_n$ where $X' \subseteq X$. A prefix sequence of X is defined as $X_{prefix} = z_1 z_2 .... z_k$ and suffix sequence of X is defined as $X_{suffix} = z_{k+1} z_{k+2} .... z_n$ for $X = z_1 z_2 .... z_k z_{k+1} .... z_n$, where $k = 1,2...n$.

Assume a database of web access sequence denoted as $Q_{database} = \{X_1, X_2 .... X_m\}$, in which the web access sequence is denoted as X_i. The length of the database is denoted as $m = |Q_{database}|$. The guidance of web access sequence X in $Q_{database}$ is the total count of single web access sequences in $Q_{database}$ that contains X, as expressed in Eq.(1).

$$\sup(X) = \left| \{ X_i \mid X \in X_i, X_i \in Q_{database} \} \right| \quad (1)$$

Any web access sequence X can acquire a guidance of at most one from web access sequence X_i. If $\sup(X) > MinSup$, then X is termed as sequential web access pattern, where $MinSup$ indicates the support threshold. On the contrary, if $\sup(z_i) > MinSup$, then z_i is termed as a frequent event. For the other condition, it can be termed as an infrequent event.

As mentioned earlier, Table I shows the web access sequence database. The web access sequence patterns are supported by at least three web access sequences, if $MinSup = 3$, as shown in Table II.

Table I Representation of Web Access Sequence In A Sample Database

Session ID	1	2	3	4
Web Access Sequence	i j l i k	m i m j k i k	j i j n i m	i n j i k n k

Table II Representation of Sequential web Access Patterns with support in Sample database

Pattern length	1	2	3	4
Web Access Sequence patterns with support	i:4 j:4 k:3	ii:4 ij:4 ik:3 ji:4 jk:3	iij:3 iji:4 ijk:3 jik:3	ijik:3

Construction of Pattern-tree

As mentioned earlier, the sequential access patterns are stored tightly through the development of a pattern-tree model. As a result, it matches the outcome with the present access sequence of the user. Further, it is supposed to produce more recommendation rule in a dynamic way. In fact, web access sequence patterns are handled as a group of strings over a fixed term D, which can also be considered as a group of unique access events. Moreover, an indication in the form of the

symbol is stamped in the pattern tree from D, which holds the associated support value. Instead of nodes, the edges of the tree are often labeled, yet the similar structure has not been available in either case.

To the next of constructing the pattern tree, the following process does not meet the situation to require the actual web access sequence database. In fact, the construction of pattern-tree mostly requires the scan of entire patterns of web access sequence. A valuable pseudo code of Patter-tree construction algorithm is shown in Algorithm 1. In this algorithm, CS-mine is used to mine the web access sequence patterns by the components of sequential pattern mining.

ALGORITHM 1: Pattern-Tree Construction

Input: A group of sequential web access patterns
Output: Pattern-tree
Design the empty root node N for patter-tree P
For each X
  Fix current node C mark to N
    For $i = 1:n$
      If C has child node z_i
        Operates with extreme support between X and fix C to z_i
      Else if
        Design a new child node with the support of X and fix C to new child node
      End if
    End for
Return Pattern-tree P

To perform the complexity analysis, the sequential web access patterns X is embedded into the constructed P. In fact, the complexity analysis is carried out by following two steps. The initial step inspects the matching prefix sequence X_{prefix} of X in P. Subsequently, the next step is to form a new path for the residual suffix sequences X_{suffix} of X in P, which is considered as the non-matching sequence. Let us consider the equal cost for testing the previous node and the constructed new node.

The overall cost of embedding a sequential web access pattern X is represented in Eq. (2), which holds the length n.

$$V\left(\left|X_{prefix}\right|\right)+V\left(\left|X_{prefix}\right|\right)=V\left(\left|S\right|\right)=V(n) \quad (2)$$

The cumulative cost required for embedding the entire sequential web-access patterns is similar to composing all the pattern-tree from scratch, which is expressed in Eq. (3), with length n.

$$V\left(\left|S_1\right|\right)+V\left(\left|S_2\right|\right)+.....+V\left(\left|S_1\right|\right)=V(n) \quad (3)$$

Since the advanced updates are regularly being included; the weblogs cannot be remained as static. Yet, it can speculate it to be comparatively static for a particular duration, in the case of web recommendation.

Generation of Recommendation rules

In the generation of the recommendation of rules, the processing components examine for recognizing the access path that is best matching in the pattern-tree. This path is selected based on the present access sequence of the user. In fact, there may be a condition of having less opportunity to detect the matching path from the pattern-tree if the present access sequence of the user seems to be longer. The appropriateness of producing the recommendation rules is enhanced by focussing on the present access sequence for the condition of not get matched with the entire access sequence. Accordingly, the matching path is determined on the basis of similar access sequence by eliminating the starting item frequently till the detection of the matching path or raising a situation of not ignoring any of the items from the access sequence. Furthermore, in the constructed patter-tree, the overall length is considered as the overall extension of the pattern tree. However, when the length of the present access sequence is not greater than the extension of the pattern-tree, then it is unable to recognize the matching path. Thus, in order to form the present access sequence lower than the overall extension of the pattern tree, it may ignore most of the starting items. The pseudo code used for the generation of recommendation roles is shown in Algorithm 2.

ALGORITHM 2: Generation of Recommendation rules

Input: Pattern-tree P
  Present Access sequence X
  Minimum length of the access sequence denoted as $MinLength$
  Maximum length of the access sequence denoted as $MaxLength$
Output: Recommendation rules Z
Initialize $Z = \Phi$
If $|X| > MaxLength$
  Eliminate the first $|X| - MaxLength + 1$ elements from X
If $|X| < MinLength$
  Return Z
Else if
  Fix C to N of P
End if
For each item in X
  If C has a child node
    Fix C point to the child node
  Else if
    Ignore the starting item from X repeat the steps
  If C has a child node
    Introduce the child nodes into Z arranged by their support
  End if
  End if
Return Z

In general, only lower accuracy can be obtained with the recommendation rules produced from the shorter matching paths. Moreover, it is essential to process the higher length web access sequence. On the contrary, it may raise the situation to conclude the matching process of sequence, if the residual access sequence is lesser than threshold.

Performance Evaluation

Performance Measures

Consider the web access sequence $X = z_1 z_2 ....z_k z_{k+1} ....z_n$. Here, $X_{prefix} = z_1 z_2 ....z_k$ be the prefix sequence and $Z = z_1, z_2, ...z_m$ be

the generated recommendation rule from pattern-tree. Moreover, the recommendation rule is assigned as correct, if $z_{k+1} \in Z$ and is incorrect for the reverse condition. In addition, the recommendation rule is assigned as m-step satisfactory if $z_i \in Z$ and is m-step unsatisfactory for the reverse condition. The measures evaluated in the experimentation are precision, satisfaction, and applicability and the appropriate definitions are portrayed as follows.

Precision: The precision of the generated recommendation rules can be represented in Eq. (4), where Z_c indicates the subset of N that comprises of right recommendation rules.

$$A = \frac{|Z_c|}{|Z|} \quad (4)$$

Satisfaction: The m-step satisfaction of the generated recommendation rules is expressed in Eq. (5), where Z_s indicates the subset of N that comprises of m-step satisfactory rules.

$$S = \frac{|Z_s|}{|Z|} \quad (5)$$

Applicability: The applicability of the web recommendation is defined in Eq. (6), where Z_n indicates the subset of N that comprises of non-empty recommendation rules.

$$B = \frac{|Z_n|}{|Z|}$$

Experimental Results

The convergence analysis of scalability with respect to support threshold is shown in Fig. 2. During the experiment, the scalability of the sequential pattern mining and patter-tree construction is measured, with respect to the support threshold is observed. Here, the runtime is measured to validate the scalability of the proposed method. Through the performance analysis, the runtime decreased with respect to the increase of support threshold.

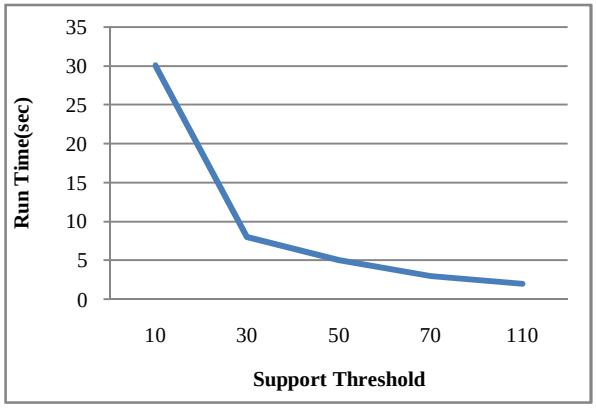

Fig 2 Convergence analysis of scalability with respect to support threshold

Further, the convergence analysis of scalability with respect to the number of recommended pages is shown in Fig. 3. It measures the satisfaction and precision of the proposed model.

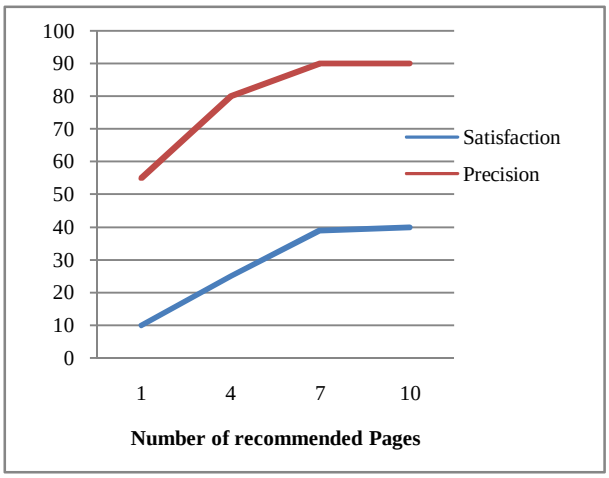

Fig 3 Convergence analysis of scalability with respect to number of recommended pages

CONCLUSION

This paper has presented the web recommendation system for information retrieval using sequential web access patterns. The proposed model mines the constant sequential web access patterns through the adoption of CS- mine. Further, the constructed pattern-tree stores the mined patterns that are useful for the subsequent matching and producing web links for online recommendations. Ultimately, the performance analysis has shown the superior performance of the proposed model.

References

1. R. Cooley, B. Mobasher, and J. Srivastava, "Data preparation for mining World Wide Web browsing patterns", *Journal of Knowledge and Information Systems*, vol. 1, no. 1. 1999.
2. R. Srikant, and R. Agrawal, "Mining sequential patterns: generalizations and performance improvements", Proc. of the 5th International Conference on Extending Database Technology (EDBT), Avignon, France, pp. 3-17, 1996.
3. D.S. Phatak, and R. Mulvaney, "Clustering for personalized mobile web usage", Proc. of the IEEE FUZZ'02, Hawaii, pp. 705- 710, May 2002.
4. B. Mobasher, "A web personalization engine based on user transaction clustering", Proc. of the 9th Workshop on Information Technologies and Systems (WITS'99), December 1999.
5. B. Mobasher, H. Dai, T. Luo and M. Nakagawa, "Effective personalization based on association rule discovery from web usage data", Proc. of the 3rd ACM Workshop on Web Information and Data Management (WIDM01), 2001.
6. T. Joachims, D. Freitag, and T. Mitchel, "WebWatcher: a tour guide for the World Wide Web", Proc. of the 5th International Joint Conference on AI, Japan, pp.. 770-775, 1997.
7. J. Konstan, B. Miller, D. Maltz, J. Herlocker, L. Gordon, and J. Riedl, "GroupLens: applying collaborative filtering to usenet news", Communications of the ACM, vol.40, no.3), pp. 77-87, 1997.

8. M. Eirinaki and M. Vazirgiannis, "Web mining for web personalization", ACM Transactions on Internet Technology, vol. 3, No. 1, 2003, pp. 1-27.
9. R. Agrawal, and R. Srikant, "Mining Sequential Patterns", Proc. of the 11th International Conference on Data Engineering, Taiwan, 1995.

A. O. Gathekar and A. M. Deshpande, "Implementation of melody extraction algorithms from polyphonic audio for Music Information Retrieval," 2016 IEEE International Conference on Advances in Electronics, Communication and Computer Technology (ICAECCT), Pune, India, 2016, pp. 6-11.

10. N. D. Ravi and D. G. Bhalke, "Musical Instrument Information retrieval using Neural Network," 2016 IEEE International Conference on Advances in Electronics, Communication and Computer Technology (ICAECCT), Pune, India, 2016, pp. 418-422.
11. H. Wang, Q. Zhang and J. Yuan, "Semantically Enhanced Medical Information Retrieval System: A Tensor Factorization Based Approach," IEEE Access, vol. 5, no. , pp. 7584-7593, 2017.
12. L. Li; Q. Xu; T. Gan; C. Tan; J. H. Lim, "A Probabilistic Model of Social Working Memory for Information Retrieval in Social Interactions," in IEEE Transactions on Cybernetics , vol.PP, no.99, pp.1-13.
13. S. Abdul-Rauf, H. Schwenk, P. Lambert and M. Nawaz, "Empirical Use of Information Retrieval to Build Synthetic Data for SMT Domain Adaptation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 4, pp. 745-754, April 2016.
14. F. Raposo, R. Ribeiro and D. Martins de Matos, "Using Generic Summarization to Improve Music Information Retrieval Tasks," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 6, pp. 1119-1128, June 2016.
15. F. Xu, Y. Q. Jin and A. Moreira, "A Preliminary Study on SAR Advanced Information Retrieval and Scene Reconstruction," in IEEE Geoscience and Remote Sensing Letters, vol. 13, no. 10, pp. 1443-1447, Oct. 2016.
16. L. Weng, L. Amsaleg, A. Morton and S. Marchand-Maillet, "A Privacy-Preserving Framework for Large-Scale Content-Based Information Retrieval," in IEEE Transactions on Information Forensics and Security, vol. 10, no. 1, pp. 152-167, Jan. 2015.
17. G. Fanti and K. Ramchandran, "Efficient Private Information Retrieval Over Unsynchronized Databases," in IEEE *Journal of Selected Topics in Signal Processing*, vol. 9, no. 7, pp. 1229-1239, Oct. 2015.
18. E. Okhapkina, V. Okhapkin and O. Kazarin, "Adaptation of Information Retrieval Methods for Identifying of Destructive Informational Influence in Social Networks," 2017 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA), Taipei, Taiwan, 2017, pp. 87-92.
19. O. Banouar and S. Raghay, "Personalized information retrieval through alignment of ontologies: State of art," 2017 8th International Conference on Information and Communication Systems (ICICS), Irbid, 2017, pp. 153-15
20. E. Pinho and C. Costa, "Extensible Architecture for Multimodal Information Retrieval in Medical Imaging Archives," 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Naples, 2016, pp. 316-322.
21. H. Sun and S. A. Jafar, "The capacity of private information retrieval with colluding databases," 2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Washington, DC, 2016, pp. 941-946.
22. B. Shah and J. Pareek, "Query Optimization for Information Retrieval in Multilingual Environment for E-governance resources," 2016 International Conference on ICT in Business Industry & Government (ICTBIG), Indore, 2016, pp. 1-4.
23. X. Ye, H. Shen, X. Ma, R. Bunescu and C. Liu, "From Word Embeddings to Document Similarities for Improved Information Retrieval in Software Engineering," 2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE), Austin, TX, 2016, pp. 404-415.
24. Chandurkar and A. Bansal, "Information Retrieval from a Structured KnowledgeBase," 2017 IEEE 11th International Conference on Semantic Computing (ICSC), San Diego, CA, 2017, pp. 407-412.
25. K. Sankar and G. N. K. S. Babu, "Web information retrievals: An excellent image portal with automated hidden tag to image," 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, 2016, pp. 150-154.
26. M. M. Rahman and C. K. Roy, "STRICT: Information retrieval based search term identification for concept location," 2017 IEEE 24th International Conference on Software Analysis, Evolution and Reengineering (SANER), Klagenfurt, pp. 79-90, 2017.
27. N. Gaur and A. Singh, "Recommender system - Making lifestyle healthy using information retrieval," 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), pp. 479-484, Dehradun, 2016.

How to cite this article:

Poonam Yadav.2017, A Method Based on Sequential Access Patterns For Information Retrieval. *Int J Recent Sci Res.* 8(7), pp. 18712-18716. DOI: http://dx.doi.org/10.24327/ijrsr.2017.0807.0554

****