Volume 3, Issue 7, July 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

New Approach to Clusters Web Sessions Based on Bio-Informatics Distance Metric Technique for Prediction Model of Web Caching and Prefetching

Dharmendra Patel*
Smt.Chandaben Mohanbhai PatelInstitute of Computer Applications, CHARUSAT,CHANGA,GUJARAT, India

Dr.Kalpesh Parikh
Director, Intellisense IT, Ahmedabad, GUJARAT, India

Abstract— A server raw log file contains much valuable information related to internet user transactions. To generate meaningful and hidden patterns from it requires mining techniques related to web objects perspective. Mining techniques from web objects perspective are divided into three main approaches: (1) Classification (2) Association Rule Mining and (3) Clustering. Raw log file transactions could not be classified into predefined classes so classification approach is not suitable for generating patterns from raw log file. Research shows that association rule mining generates many unnecessary rules so analysis of pattern is very difficult. Clustering approach of web mining is an ideal solution to generate meaningful patterns from raw log file. Clustering techniques are dividing into two parts: (1) Hierarchical and (2) Partitioning relocation clustering. Hierarchical clustering is not an efficient because most of all algorithms do not revisit clusters once constructed while partitioning clustering provides improvement based on relocation. Clustering approach builds clusters based on distance metric and generation of distance metric from string sequences is very complex task. This paper studies different approaches to build distance metric. This paper uses well known bio informatics algorithms that are used for protein sequences similarity to generate distance metric and compares it with well known edit distance algorithm. This paper also provides a new approach for formation of clusters in context of prediction model of web caching and prefetching.

Keywords— Web Caching, Web Prefetching, Data Mining, Clustering, Association rule Mining, Distance Metric, Fuzzy C-Means

I. INTRODUCTION

Server raw log files consist of much useful and hidden information about internet users that are essential for web object managements for web caching and prefetching perspective. Data mining [1] is one approach for web prefetching for optimization of web objects so access latency could be reduced. Data mining techniques are divided into three main categories of algorithms to generate hidden and useful pattern from any kind of data [2]. Three categories of data mining are: (1) Classification (2) Association Rule Mining (3) Clustering. Classification approach is out of scope of this research because predefined classes could not be formed form raw log file. Several researches have been done in association rule mining technique to generate useful and hidden information from large set of data. Author [3] describes that association rule mining discovers homogeneous pages that are commonly accessed together in same session. In research [4] author identifies an advantage of association rule mining that association rule mining provides interesting relationship among items of large data set but it produces too many useless rules results in inconsistent prediction so association rule mining is not fitted in web prefetching research. Clustering is best technique for web prefetching purpose because it improves efficiency and scalability of real applications tasks [5, 6]. Clustering is unsupervised learning technique that is very useful for web applications [7, 8]. Clustering algorithms are classified into many classifications: (1) Hierarchical Clustering (2) Partitioning methods and (3) Grid based methods. Researches show that most of hierarchical algorithms do not revisit clusters once they build so this technique is not efficient for clusters formation. Partitioning clustering technique has different relocation schemes so optimization in cluster formation is achieved but is not suitable for categorical data and it requires mentioning k clusters for formation of sub clusters. Relocation concept of partitioning clustering algorithms is good for formation of clusters but it only favoured numerical data so in proposed research this approach is also not fitted as a whole. According to density based algorithms space is divided into set of connected components and based on density function cluster formation is to be done. Generally such kinds of algorithms are useful for spatial data so this technique is also out of scope of proposed research. Literature survey showed that there is no efficient technique for cluster formation for web objects so new approach is formed based on distance metric. Distance metric is useful to determine similarity among objects by means of specific technique that transforms non numerical attributes to numerical category. In [9] author discussed the role of distance metric as a similarity measure and described data distribution that is not suitable for Euclidean distance. In study [10] author discussed well known Levenshtein edit distance measure for string similarity which requires minimum numbers of insertions, deletions and modifications to transform one string to another. In bio informatics field, distance measurement is a vital stage to compare DNA sequences. In study [11], author uses bioinformatics algorithms that use edit distance and hamming distance for acceleration of biological sequences. Needleman-Wunsch algorithm [12] and Smith-Waterman algorithm [13] are two widely used approaches in bio informatics to provide similarity among DNA sequences. Needleman-Wunsch algorithms

has a global alignment problem that measures the similarity between two string sequences while Smith-Waterman finds the longest similarity within two string sequences. This paper provides new approach to clusters web sessions using distance metric and to generate distance metric both edit distance and combination of edit and hamming code distance techniques are to be used. This paper also analyze how bio informatics distance metric approach that uses both edit and hamming distance is an appropriate way for formation of clusters. This paper uses new technique for formation of clusters which is similar as Fuzzy C-Means but that is based on distance threshold and numbers of clusters are not fixed.

Section-2 of paper will discuss different well known techniques for generating distance metric. Section-3 will discuss approach for formation of clusters. Section-4 will provide conclusion of paper.

II. DISTANCE METRIC TECHNIQUES

In mathematics metric is used as a measure of distance so any calculation related to distance is stored in distance metric. Distance metric is an efficient tool for clusters formation. Euclidean distance considers geometric distance between two high dimensional data objects and it is not suitable for string similarity. Edit distance like Levenshtein is a string metric for measuring the difference between two strings sequences so this kind of measure is suitable for formation of clusters for web sessions because web sessions are generally a string sequence of number of web objects. According to Levensthein distance, distance between two words of string is minimum numbers of insertions, deletions or substitutions required to change one word of string to other. There are other popular techniques for edit distance:

(1) Longest common subsequences (LCS):- It is used to find longest subsequence of string common to all sequences of strings. It only allows insertion and deletion edits.

(2) Hamming Distance (HD): - It is applicable to only similar length strings and allows only substitution edits.

Several bio informatics techniques are available that use combination of edit as well as hamming distance. Two most popular techniques of this category are: (1) Needleman-Wunsch and (2) Smith-Waterman. Needleman-Wunsch algorithm is used in bioinformatics for the alignments of protein sequences. Similarity Matrix is used by this technique to assign scores for aligned characters. It is a global alignment technique means closely related sequences of same length are appropriate for it. Smith- Waterman technique is based on local alignment which determines similar regions between two strings of any sequence. Following table shows characteristics of different distance metric techniques.

Table-1 Distance Metric Techniques

Sr.No	Technique	Description	Advantages	Disadvantages
1.	Euclidean Distance	It describes distance between two points that would be measure with ruler and calculated using Pythagorean theorem.	(1)It is faster for determination of correlation among points (2) It is fair measure because it compares data points based on actual ratings.	(1)It is not suitable for ordinal data like string. (2) It requires actual data not rank.
2.	Levenshtein	It is a string metric for measuring the difference between two strings.	It is fast and best suited for strings similarity.	It is not considered order of sequence of characters while comparing.
3.	Needleman-Wunsch	It is a bio informatics algorithm and provides global alignment between strings while comparing.	It is best for string comparison because it considers ordering of sequence of characters	It requires same length of string while comparing.
4.	Smith-Waterman	It is a bio informatics algorithm and provides local alignment between strings while comparing.	It is best for string comparison because it considers ordering of sequence of characters and it is applicable for either similar or dissimilar length of strings.	It is quite complex than any global alignment technique.

From above table it is to be analysed that Euclidean distance is not suitable for proposed research because web sessions consists of sequences of web objects and which are in string format. Levenshtein distance is a very good technique for string sequences similarity but for prediction model of web caching and prefetching an ordering of web objects is an important aspect that is ignored by this distance metric technique so it is also not an appropriate way in proposed research context. Both Needleman-Wunsch and Smith-Waterman considers an ordering of sequence for string

matching so they are ideal for this context. Web Sessions are not always of same length so Needleman-Wunsch algorithm is not cent percent fit for formation of web sessions clusters as it only provides global alignment. Smith-Waterman algorithm is applicable for both same length sequence as well as dissimilar length of sequences so it is an ideal algorithm for formation of clusters in this proposed research.

III. APPROACH TO FORMATION OF CLUSTERS

As discussed in introduction part hierarchical way of clustering is not an efficient technique for formation of clusters because clusters are not revisited by most of algorithms of hierarchical technique. For the prediction model of web caching and prefetching partitioning relocation clustering is an ideal solution for clusters formations. There are two main techniques for partitioning clustering and they are: (1) K-means and (2) K-medoids. In both techniques specification of **K** is require and it indicates numbers of clusters that are going to form after implementation of that technique. In proposed research it is not able to predict number of clusters so it is impossible to give value of **K** at initial level so both of these techniques are not suitable for this context. In both k-means and k-medoids one object could not be the part of more than one cluster while in proposed research context, one web object could be part of more than one clusters. One popular technique of clustering is Fuzzy C-Means that attempts to divide any **n** elements into collections of **m** fuzzy clusters with some criterion but this algorithm also requires choosing a number of clusters so it is not fitted perfectly in this proposed research work.

One new approach for formation of clusters is suggested in this proposed work. Following are number of steps of this approach:

[1] Determine distance metric based on smith-waterman distance metric technique.
[2] Decide threshold value in context of proxy server cache memory.
[3] Based on threshold value form clusters of web objects which fit according to threshold value.
[4] Repeat step 3 based on new threshold value if require.

This new approach des not require to specify **K** in advance and other thing is one object may be part of more than one cluster so fuzzy criterion is also involved in it.

IV. CONCLUSIONS

Clustering is an important technique to determine patterns from raw log server file because it is very simple to implement and does not require predefined classes. In context to formation of web sessions clusters, it is vey challenging task because web sessions are sequences of strings and to determine string similarity is most crucial task. In this paper some techniques for distance metric generation is discussed and also analysis of all techniques in context to string similarity is done. Analysis of string similarity distance metric approaches identifies Needleman-Wunsch and Smith-Waterman techniques relevant to proposed work. Finally paper identified best suited technique in context to research work that could be implemented for any kind of string whether similar or dissimilar. The paper also discussed several ways in formation of clusters after distance metric generation. Paper identified that partitioning clustering approaches are best for formation of clusters but algorithms related to it require specification of numbers of clusters in advance and one object could not be part of other cluster so they are not fitted in proposed research. Paper discussed fuzzy C-means approach for cluster formation because it might include one object in more than one cluster but still it requires specifying number of clusters in advance so it could not be implemented as it is in proposed research. At last paper discussed new approach for web sessions clustering in context of prediction model of web caching and prefetching that is an ideal way for formation of clusters.

REFERENCES

[1] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.
[2] M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective. IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996
[3] J. Pitkow and P. Pirolli, "Mining longest Repeating subsequences to Predict World Wide Web Surfing", *Proceedings USENIX Symposium on Internet Technologies and Systems(USITS'99),*(1999).
[4] Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proc. of the 1993 ACM SIGMOD International Conference on Management of Data SIGMOD 1993, Washington, DC, pp. 207–216 (1993).
[5] F. Khalil, J. LiAn, and H. Wang, "Integrated model for next page access prediction", *Int. J. Knowledge and Web Intelligence,* 1(1-2), (2009), pp. 48-80(33)
[6] M. Rigou, S. Sirmakesses, and G. Tzimas, "A method for personalized clustering in data intensive web applications", *APS'06, Denmark,* (2006), pp. 35–40.
[7] COOLEY, R., MOBASHER, B., and SRIVASTAVA, J. 1999. Data preparation for mining world wide web browsing. Journal of Knowledge Information Systems, 1, 1, 5-32.
[8] HEER, J. and CHI, E. 2001. Identification of Web user traffic composition using multimodal clustering and information scent. *1st SIAM ICDM, Workshop on Web Mining*, 51- 58, Chicago, IL.
[9] Jie Yu, Amroes J , Sebe N, Qi Tian , A new study on distance metric as similarity measurement, Multimedia and Expo, IEEE International conference July, 2006 , pages- 533-536.
[10] Ristad E.S, Yianilos P.N, Learning string edit distance, Pattern analysis and machine intelligence, IEEE Transactions, May 1998, volume 20, Issue-5, pages-522-532.

[11] Zubair Nawaz, Mudassir Shabbir, Zaid Al-Ars, Koen Bertels, Acceleration of Biological Sequence Alignment using Recursive Variable Expansion, pages-233-237.

[12] Saul B. Needleman, Christian D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, Journal of Molecular Biology 48 (3): 443-53, 1970.

[13] Temple F. Smith, Michael S. Waterman, Identification of Common molecular Subsequences, Journal of Molecular Biology 147: 195-197,1981.

