STUDY OF DISTANCE MEASUREMENT TECHNIQUES IN CONTEXT TO PREDICTION MODEL OF WEB CACHING AND WEB PREFETCHING

Dharmendra Patel

[1]Smt.Chadaben Mohanbhai Patel Institute of Computer Applications, Charotar University of Science and Technology, Changa, Gujarat, India

ABSTRACT

Internet is the boon in modern era as every organization uses it for dissemination of information and e-commerce related applications. Sometimes people of organization feel delay while accessing internet in spite of proper bandwidth. Prediction model of web caching and prefetching is an ideal solution of this delay problem. Prediction model analysing history of internet user from server raw log files and determine future sequence of web objects and placed all web objects to nearer to the user so access latency could be reduced to some extent and problem of delay is to be solved. To determine sequence of future web objects, it is necessary to determine proximity of one web object with other by identifying proper distance metric technique related to web caching and prefetching. This paper studies different distance metric techniques and concludes that bio informatics based distance metric techniques are ideal in context to Web Caching and Web Prefetching.

KEYWORDS

Web Caching, Web Prefetching, Distance Metric, Access Latency, Bio Informatics

1. INTRODUCTION

WWW is an information hub that consists of enormous web objects in form of web page, audio, video, image etc. The common problem for internet user is delay while accessing web objects in spite of proper bandwidth. Web Caching and Web Prefetching concepts integration [1] is the software solution of delay problem. The main purpose of Web Prefetching is to generate patterns from historical data. Following figure describes main approaches of pattern discovery.

Dependency Graph It increases network traffic and prediction accuracy is very low [2]. Markov Model is well known model that predicts the next page by matching user's current access sequence with the user's historical web access sequence but main limitations are; In the lowest order Markov model, the prediction is not accurate while in high order Markov model converge is low and complexity is more [3].Cost Function is less popular and it uses popularity and lifetime kind of measures to generate pattern [4]. New pattern generation is quite impossible and very complex in all techniques of dependency graph, Markov Model and Cost Function. Data Mining [5] is the ultimate technique of web Prefetching that generates new patterns from large data set. The techniques of data mining applied for web data are known as Web Mining [6]. Web mining has three main categories (i) Web Content (ii) Web Structure and (iii) Web Usage. Web

DOI :10.5121/ijscai.2016.5101

Caching and Prefetching based prediction model uses the data of server raw log file to generate meaningful sequence of patterns means Web Usage Mining category of Web mining is applicable for this. Web Usage mining has three main steps (a) Data Preprocessing (b) Pattern Discovery and (c) Pattern Analysis[7]. Number of researches have been done on data preprocessing stage [8][9][10][11][12][13] but very few researches have been done for pattern discovery and analysis. This paper focuses on pattern discovery and analysis in exhaustive way. For pattern discovery and analysis stage it is required to identify sessions of particular user.

There are number of sessionization techniques [14][15][16] are identified in literature and technique of combination of IP,agent and sessionization heuristic is an ideal in context to prediction model as no additional overhead to determine sessions as well as data regarding IP and agent are always available. Once sessions are formed it is challenge to determine proximity of one session with another in order to generate interesting, new, hidden and valuable sequence of pattern in the context to user. To determine the proximity among objects number of distance metric techniques [17][18][19] are identified in literature. The detail study and identification of an appropriate distance measurement is a vital for generation of sequence of patterns. The section 2 will describe detail study of distance metric techniques in context to web caching and Prefetching. The Section 3 provides conclusion about all distance measurement techniques in context to Web Caching and Web Prefetching.

2. STUDY OF DISTANCE MEASUREMENT TECHNIQUES

Most popular and commonly used distance metric technique studied in literature is Euclidean distance [20,21, 22]. It is an ordinary distance between two points that is measured with the ruler and it is derived from Pythagorean formula. The distance between two data points in the plane with coordinates (p, q) and (r, s) is formulated by:

$$\textbf{DIST ((p, q), (r, s)) = Sqrt ((p-r)}^{2} + \textbf{(q-s)}^{2}\textbf{)}$$

The usefulness of Euclidean depends on circumstances. For any plane, it provides pretty good results, but with slow speed. It is also not useful in the case of string distance measurement. Euclidean distance can be extended to any dimensions. Another most popular distance metric technique is Manhattan distance that computes the distance from one data point to another if grid like a path is followed. The Manhattan distance between two items is the sum of the differences of their corresponding elements [23]. The distance between two data points with coordinates (p1, q1) and (p2, q2) is derived from:

$$\textbf{DIST ((p1, q1), (p2, q2)) = } \sum_{i=1}^{n} | pi - qi |$$

It is the summation of horizontal and vertical elements where diagonal distance is computed using Pythagorean formula. It is derived from the Euclidean distance so it exhibits similar characteristics as of Euclidean distance. It is generally useful in gaming applications like chess to determine diagonal distance from one element to another. Minkowski is another famous distance measurement technique that can be considered as a generalization of both Euclidean and Manhattan distance. Several researches [24,25,26] have been done based on Minkowski distance measurement technique to determine similarity among objects. The Minkowski distance of order *p* between two points: P (x1, x2, x3…xn) and Q(y1,y2,y3,…yn) R^n is defined as:

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}.$$

If the value of p equals to one $_{s1, s2}$ Minkowski distance is known as Manhattan distance while for value 2 it becomes Euclidean distance. All distances discussed so far are considered as ordinary distances that are measured from ruler and they are not suitable for measuring similarity of two strings so they are not appropriate in context to propose research since web session consists of numbers of string in the form of URLs. Hamming distance is a most popular similarity measure for strings that determines similarity by considering the number of positions at which the corresponding characters are different. More formally hamming distance between two strings P and Q is $|\ Pi\ \acute{H}\ Qi\ |$. Hamming distance theory is used widely in several applications like quantification of information, study of the properties of code and for secure communication. The main limitation on Hamming distance, is it only applicable for strings of the same length.

Hamming distance is widely used in error free communication [27,28] field because the byte length remains same for both parties who are involved in communication. Hamming distance is not so much useful in context to web usage mining since length of all web sessions is not similar. The Levenshtein distance or edit distance is more sophisticated distance measurement technique for string similarity. It is the key distance in several fields such as optical character recognition, text processing, computational biology, fraud detection, cryptography etc. and was studied extensively by many authors [29,30,31]. The formula of Levenshtein distance between two strings S1, S2 is given by **Lev** $_{\mathbf{s1, s2}}$ (| S1|, |S2|) where

Lev $_{\mathbf{S1, S2}}$ (I, j) = Max (I, j) if Min (I, j) =0,
Otherwise
Lev $_{S1,S2}$ (i,j) = Min (Lev $_{s1, s2}$ (i-1, j) +1)
OR
Min (Lev $_{s1, s2}$(i, j-1) + 1)
OR
Min (Lev $_{s1, s2}$ (i-1, j-1) + [S1i # S2j]

Levenshtein distance measurement technique is an ideal context for web session since it is applicable to strings of unequal size.

Several bioinformatics distance measurement techniques that are used to align protein or nucleotide sequences can be used to web mining perspectives to cluster unequal size web sessions. One of the most important techniques of this category was invented by Saul B. Needleman and Christian D. Wunsch [32] to align unequal size protein sequences. This technique uses dynamic programming means solving complex problems by breaking them down into simpler sub problems. It is a global alignment technique in which closely related sequences of same length are very much appropriate. Alignment is done from beginning till end of sequence to find out best possible alignment. This technique uses scoring system. Positive or higher value is assigned for a match and a negative or a lower value is assigned for mismatch. It uses gap penalties to maximize the meaning of sequence. This gap penalty is subtracted from each gap that has been introduced. There are two main types of gap penalties such as open and extension. The open penalty is always applied at the start of the gap, and then the other gaps following it are given with a gap extension penalty which will be less compared to the open penalty. Typical values are –12 for gap opening, and –4 for gap extension. According to Needleman Wunsch algorithm, initial matrix is created with N * M dimension, where N = number of rows equals to number of characters of first string plus one and M= number of columns equals to number of characters of first string plus one. Extra row and column is used to align with gap. After that scoring scheme is introduced that can be user defined with specific scores. The simple basic scoring scheme is, if two sequences at i^{th} and j^{th} positions are same matching score is 1(S(I,j) =1) or if two sequences at i^{th} and j^{th} positions are not same mismatch score is assumed as -1 (S(I,j)=-

1). The gap penalty is assumed as -1. When any kind of operation is performed like insertion or deletion, the dynamic programming matrix is defined with three different steps:

1. Initialization Phase: - In initialization phase the gap score can be added to previous cell of the row and column.

2. Matrix Filling Phase: - It is most crucial phase and matrix filling starting from the upper left hand corner of the matrix. It is required to know the diagonal, left and right score of the current position in order to find maximum score of each cell. From the assumed values, add match or mismatch score to diagonal value. Same way repeat the process for left and right value. Take the maximum of three values (i.e. diagonal, right and left) and fill ith and jth positions with obtained score. The equation to calculate the maximum score is as under:

$$Mi,j = \text{Max}\ [\ M_{i-1,j-1} + S_{i,j},\ M_{i,j-1} + W,\ M_{i-1,j} + W]$$

Where i,j describes row and columns.
M is the matrix value of the required cell (stated as $M_{i,j}$)
S is the score of the required cell ($S_{i,\ j}$)
W is the gap alignment

3. **Alignment through Trace Back: -** It is the final step in Needleman Wunsch algorithm that is trace back for the best possible alignment. The best alignment among the alignments can be identified by using the maximum alignment score.

Needleman Wunsch distance measurement technique is an ideal one in string similarity so this technique is also considered in the proposed research context.

Smith-waterman is an important bioinformatics technique to align different strings. This technique compares segments of all possible lengths and optimizes the measure of similarity. Temple F.Smith and Michael S.Waterman [33] were founders of this technique. The main difference in the comparison of Needleman Wunsh is that negative scoring matrix cells are set to zero that makes local alignment visible. This technique compares diversified length segments instead of looking at entire sequence at once. The main advantages of smith waterman technique are:

- To gives conserved regions between the two sequences.
- To align partially overlapping sequences.
- To align the subsequence of the sequence to itself.

Alike Needleman-Wunsch, this technique also uses scoring matrix system. For scoring system and gap analysis, same concepts used in Needleman Wunsch are applicable here in Smith-Waterman. It also uses same steps of initialization, matrix filling and alignment through trace back. The equation to calculate maximum score is same as Needleman Wunsch. The main differences between Needleman-Wunsch and Smith Waterman are:

- Needleman Wunsch does global alignment while Smith Waterman focuses on local alignment.
- Needleman Wunsch requires alignment score for pair of remainder to be >=0 while for Smith Waterman it may be positive or negative.
- For Needleman Wunsch no gap penalty is required for processing while for Smith and Waterman gap penalty is required for efficient work.

- In Needleman and Wunsch score can not be decrease between two cells of a pathway while in Smith Waterman score can increase, decrease or remain same between two cells of pathway.

Table 1 describes the comparison of different distance metrics techniques in context to proposed research.

Table 1 Comparison of distance metrics 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 analyzed 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.

3. CONCLUSION

In era of internet, to reduce access latency of user is most challenging job. To reduce access latency, Web Caching and Prefetching based approach is an ideal but success of it depends on accuracy of patterns as well as capacity of generation of new patterns. This paper described most efficient technique to generate more accurate patters. This paper dealt with distance measurement techniques in order to determine proximity among web sessions. The challenge is to identifying best technique in context to web caching and Prefetching. This paper studies in depth about distance measurement techniques and identify theoretically which techniques are suitable in context to the application of Web Caching and Prefetching.

REFERENCES

[1] Sarina Sulaiman, Siti, Ajith Abraham, Shahida Sulaiman, “Web Caching and Prefetching: What, Why, and How? “, IEEE, Information Technology,2008,Volume-4,pages: 1-8.

[2] A. Nanopoulos, D. Katsaros, and Y. Manolopoulos “ A data mining algorithm for generalized Web Prefetching” IEEE Trans on Knowledge and Data Engineering , 15 (5) ,(2003). PP. 11552-1169.

[3] Dharmendra T.Patel, Dr.Kalpesh Parikh, “Quantitative Study of Markov Model for Prediction of User Behavior for Web Caching and Prefetching Purpose”, International Journal of Computer Applications (0975- 8887) , Volume 65 No.15 , March 2013,PP. 39-49.

[4] E. P. Markatos and C. E. Chronaki : A Top-10 approach to prefetching on the Web “, Proceedings of INET'98 Geneva, Switzerland, (1998), pp. 276-290.

[5] Xindong Wu, Vipin Kumar, J. Ross Quinlan, “Top 10 Algorithms in Data Mining”, Springer, Knowledge Information System,2008,PP. 1-37.

[6] COOLEY, R., MOBASHER, B. ; SRIVASTAVA, J.,” WEB MINING: INFORMATION AND PATTERN DISCOVERY ON THE WORLD WIDE WEB “,TOOLS WITH ARTIFICIAL INTELLIGENCE, 1997. PROCEEDINGS., NINTH IEEE INTERNATIONAL CONFERENCE ON 3-8 NOV 1997,PP.-558-567.

[7] Xinjin Li, Sujing Zhang,” Application of Web Usage Mining in e-learning Platform”, IEE 2010 International Conference on E-Business and E-Government, May 2010,PP-1391-1394.

[8] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan,” Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data”, SIGKDD Explorations,2000, Vol.1(2) ,PP 1-12.

[9] Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd, Hafizul Fahri Hanafi, Mohamad Farhan Mohamad Mohsin- “ Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm”, World Academy of Science, Engineering and Technology 48 2008.

[10] Tanasa D., Trousse B.. Advanced data preprocessing for intersites Web usage mining. IntelligentSystems, IEEE,2004(19), PP 59 – 65.

[11] Yuan, F., L.-J. Wang, et al.,” Study on Data Preprocessing Algorithm in Web Log Mining”, Proceedings of the Second International Conference on Machine Learning and Cybernetics, Wan, 2-5 November 2003.

[12] Pabarskaite, Z,” Implementing Advanced Cleaning and End-User Interpretability Technologies in Web Log Mining”. 24th Int. Conf. information Technology Interfaces /TI 2002, June 24-27, 2002, Cavtat, Croatia.

[13] Li Chaofeng, “ Research and Development of Data Preprocessing in Web Usage Mining”, International Conference on Management Science and Engineering , 2006.

[14] Robert.Cooley,Bamshed Mobasher, and Jaideep Srinivastava, “ Web mining:Information and Pattern Discovery on the World Wide Web”, In International conference on Tools with Artificial Intelligence, pages 558-567, Newport Beach, IEEE,1997.

[15] Robert F.Dell ,Pablo E.Roman, and Juan D.Velasquez, “Web User Session Reconstruction Using Integer Programming,” , IEEE/ACM International Conference on Web Intelligence and Intelligent Agent,2008.

[16] Spilipoulou M.and Mobasher B, Berendt B.,”A framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis”, INFORMS Journal on Computing Spring ,2003.

[17] Frank Y. Shih, Yi-Ta Wu,” Fast Euclidean distance transformation in two scans using a 3 3 neighborhood”, Computer Vision and Image Understanding ,Elsveir,2004,PP 195–205.

[18] L. Kari, S. Konstantinidis, S. Perron, G. Wozniak, J. Xu, “Computing the Hamming distance of a regular language in quadratic time,” WSEAS Transactions on Information Science & Applications,vol. 1, pp. 445–449, 2004.

[19] Stavros Konstantinidis, “Computing the Levenshtein Distance of a Regular Language”, In the Proc. of IEEE ISOC ITW2005 on Coding and Complexity; editor M.J. Dinneen; co-chairs U. Speidel and D. Taylor; PP 113-116.

[20] A. Alfakih, A. Khandani, and H. Wolkowicz ”Solving Euclidean distance matrix completion problems via semidenite programming”, Comput. Optim. (1999), PP.13-30.

[21] M. Bakonyi and C. R. Johnson, “ The Euclidean distance matrix completion problem”, SIAM J. Matrix Anal. (1995), PP 645-654.

[22] H. X. Huang, Z. A. Liang, and P. M Pardalos,”Some properties for the Euclidean distancematrix and positive semidenite matrix completion problems”, J. Global Optim.,(2003),PP 3-21.

[23] Sung-Hyuk Cha, “ Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions”, International Journal of Mathematical Models and Methods in Applied Sciences, Issue 4, Volume 1, (2007).

[24] Agrawal P. K., Flato E., Halperin D. : Polygon decomposition for efficient construction of minkowski sums. Comput. Geom. Theory Appl. 21 , 1 (2002), 39–61.

[25] Foglel E., Halperin D. Exact and efficient construction of minkowski sums of convex polyhedra with applications. In Proc. ALENEX 2006 (2006).

[26] G Ritzmann P., Sturmfels B. : Minkowski addition of polytopes: Computational complexity and applications to grobner basis. SIAM J. Discrete Math. 6 , 2 (1993), 246–269.

[27] A. Ambainis, W. Gasarch, A. Srinavasan, A. Utis: Lower bounds on the deterministic and quantum communication complexity of hamming distance, cs.CC/0411076, (2004).

[28] Wei Huang, Yaoyun Shi, Shengyu Zhang, Yufan Zhu: The coomunication complexity of Hamming distance problem, Elsevier, Information Processing Letters, (2006), pp-149-153.

[29] Navarro. A: guided tour to approximate string matching, ACM Comput. Surv. , 33(1):31–88, (2001).

[30] Andoni and R. Krauthgamer: The computational hardness of estimating edit distance. In Proceedings of the Symposium on Foundations of Computer Science, (2007).

[31] G. Navarro, R. Baeza-Yates, E. Sutinen, and J. Tarhio: Indexing methods for approximate string matching, IEEE Data Engineering Bulletin, 24(4):19–27, (2001). Special issue on Text and Databases Invited paper.

[32] Needleman Saul B and Wunsch Christian D. :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.

[33] Smith, Temple F.; and Waterman, Michael S. : Identification of Common Molecular Subsequences, Journal of Molecular Biology , pages 195–197.

AUTHOR

He received his Ph.D degree from Kadi Sarva Vishwavidyalaya, Gandhinagar in Web Mining field. Currently he is working as an associate professor at Smt.Chandaben Mohanbhai Patel Institute of Computer Applications,CHARUSAT,Changa,Gujarat,India. He has published 12 research papers in international journals of repute. He is associated with many journals of repute as an editorial board/reviewer board member. He is member of several professional bodies.