Web site audience segmentation using hybrid alignment techniques

Vinh-Trung Luu, Germain Forestier, Frédéric Fondement, Pierre-Alain Muller

MIPS, Université de Haute Alsace,
12, rue des frères Lumière - 68093 Mulhouse cedex - France
{trung.luu-vinh, germain.forestier,
frederic.fondement,pierre-alain.muller}@uha.fr

Abstract. We are working on behavioral marketing in the Internet. On one hand we observe the behavior of visitors, and on the other hand we trigger (in real-time) stimulations intended to alter this behavior. Real-time and mass-customization are the two challenges that we have to address. In this paper, we present a hybrid approach for clustering visitor sessions, based on a combination of global and local sequence alignments, such as Needleman-Wunsch and Smith-Waterman. Our goal is to define very simple approaches able to address about 80% of visitor sessions to be segmented, and which can be easily turned into small pieces of program, to be run in parallel in thousands of web browsers.

Keywords: web mining, sequential pattern mining, clustering

1 Introduction

Behavioral marketing in the Internet includes adapting web sites to the interests of the visitors in real-time, while they are browsing. Web usage mining has been widely used to transform low-level browsing data (such as page- and click-stream) into actionable knowledge, which makes sense in the business arena. This calls for operators able to compute a measure of similarity between any two sessions, in order to define groups of similar sessions, and further to segment the audience. In our case, as we want to act in real-time, we also have to provide similarity operators which can be executed quickly (in a time which is compatible with the browsing speed of visitors). Sessions can be considered as sequences of events. The granularity of these events can be fine-tuned, from pages-loads down to low-level JavaScript events. In this paper, for the sake of simplicity, we will talk of sequences of symbols, such as A-B-C-D-E-F. Luckily, we have access daily to hundred thousands of such sequences, which are recorded by our industrial partner (BeamPulse). These sequences originate mainly from e-commerce applications.

There is a large amount experience in sequence analysis in the field of DNA sequences comparison. Sequence alignment has been widely used to identify regions of similarity of DNA, RNA or protein sequences in bioinformatics. Two main approaches - global and local alignment of sequences - have been proposed, respectively by Needleman-Wunsch[1] (NW) and Smith-Waterman[2] (SW). Weinan

Wang et al.[3] labeled sitemaps as tree structures and compared pairs of sessions using sequences comparison by applying global sequence alignment. In another research, LI Chaofeng et al.[4] introduced a scoring method by combination of visiting time and URLs similar to [3]. However, according to Poornalatha Ga et al.[5], the optimal similarity between two sequences or clustering outliers can be found by an algorithm based on Smith-Waterman local alignment method only. An approach to cluster web sessions was proposed by Bhupendra S. Chordia et al. [6] where the clusters are initialized using the longest and most dissimilar sequences. Then, a combination of local and global alignment is used to update the clusters. Alternatively, Costantinos Dimopoulos et al.[7] modeled users navigation history and web page content using weighted suffix trees. Their system was then used for the prediction of web page usage. Dynamic Time Warping (DTW), another widely used sequence alignment, has been frequently used to cluster time series [8]. For example, Warissara Meesrikamolkul et al.[9] proposed a method to combine DTW and K-Means to cluster time-series efficiently. They significantly improved the execution time and improved the accuracy of the clusters. Meanwhile, Atsuyoshi Nakamura et al.[10] studied a method named packing alignment to study sequences of various length. This method is partly similar to DTW but allows gaps and limits consecutive events. DTW approach has also been the basis or reference to propose new algorithms, such as[11] by Alice Marascu et.al taking distance measure LCSS having similar distance matrix like DTW into consideration to detect similarity matching in data streams.

This paper aim at introducing a new approach for clustering sessions, by defining a combination of local and global sequence alignments for computing similarity between two pages visit sequences. As we do not target 100% applicability, issues such as impreciseness of sequence glocal (hybrid) metric, caused by the somewhat ignorance of sequences dissimilarity when lengths are uneven, can be overcome.

The remainder of this paper is organized as follows: Section 2 details our approach and provides illustrative examples. Experimental result is described in Section 3. In Section 4, we present related works and explain how our approach compares to theses earlier techniques. Finally, section 5 concludes our paper and suggests some future research directions.

2 Proposed method

As introduced previously, the Needleman-Wunsch [1] (NW) algorithm creates a global alignment of two sequences. This algorithm aims at detecting the optimal alignment over the entire length of two sequences. Thus, this algorithm is appropriate to align pair of sequences of similar length. Meanwhile, the Smith-Waterman [2] (SW) algorithm is dedicated to local sequence alignment and is then suitable when comparing two sequences with significant difference in lengths. In this paper, we used the NW scoring scheme of +1 for matching and -1 for non-matching pair of items in sequences, and the SW scoring scheme of +2 for matching and -1 for non-matching inside matching, ignore non-matching

outside. We selected these two algorithms for their simple and efficient alignment scoring scheme. To detect the best alignment of sequences pair, these two algorithms use a matrix with a number of rows and columns corresponding to the sequences lengths. This matrix is filled by aligning score between these two sequences, and finally a trace back is performed [1, 2].

NW and SW, by their featured alignments, measure similarity of sequence pairs in different evaluations. For instance, SW score of Figure 1 and NW scores of Figure 2 and 3 are equal. However, if we take the rate of similarity lengths over sequence lengths into consideration, the similarity of Figure 1 is not as much as Figure 2 and 3.

Fig. 1. Sequence alignment on two sequences having a common subsequence but different lengths

Fig. 2. Sequence alignment on two identical sequences

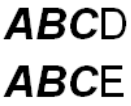

Fig. 3. Sequence alignment on two sequences having a common subsequence and similar lengths

In the comparison of web access sequences, the pairs of sequences in Figure 2 and 3 are more likely to be similar than the sequences of Figure 1 as they contains the same number of items. However, their similarity scores are mostly the same. Alternatively, Figure 4 and 5 show cases of pairs of sequences that have the same length and the same NW score. However, the first pair in Figure 4 is more consecutive than the one in Figure 5. In other words, SW score of the first pair is higher than the second. This consecution, in our opinion, makes the first pair more similar in web access sequence comparison.

Fig. 4. Sequence alignment on two sequences having a common subsequence and similar lengths

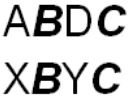

Fig. 5. Sequence alignment on two sequences having common subsequences and similar lengths

Since the clustering of web sessions is based on the alignment scores, these scores have to reflect the real similarity of the sequences. In our opinion, the *real* similarity of web sessions pair should not only consider a specific rate of common pages but should also take into account the consecution in those common pages. This consecution plays an important role in web usage mining, where the same set of pages but in different order represents dissimilar accessing behaviors. Accordingly, we propose a method to compute what is expected of similar pairs of sequences. As mentioned previously, a global view in sequence alignment is NW strong advantage. Its scoring scheme takes both similarity and dissimilarity into account but does not really reflect the consecution of similar items. Therefore, another algorithm focusing on this consecution should be employed to process the result provided by NW. SW is a good candidate as it focuses on local similarity in sequence alignment. Thus, the method proposed in this paper takes the advantages of NW and SW and reduce their disadvantages in web access sequence alignment.

We selected five pairs of sequences: (ABCDEFG, BCDEFG), (ABCDEFGH, ABXDYFGH), (ABCDEFG, CDEFG) (DEFG, DEFG) and (ABCDEF, CDEF) that have specific properties to illustrate the method. We proposed a set of rules that combine NW and SW alignment scores. We expect that similar pairs of sequences match the rules for both alignments. Furthermore, the order of the rules should not affect the final result. We recommend to first check the rules using NW alignment score and then the rules using SW alignment. Using this process, we start by considering a global alignment of the sequences and then a local alignment. In other words, SW alignment works on NWs result. We define rule matching (✓) and rule non-matching (✗) pairs through checking as result in Table 1.

Table 1. Rule matching and non-matching pairs in sequence alignments result

	NW score > 2	NW score > 2 and SW score > 10
ABCDEFG BCDEFG	✓	✓
ABCDEFGH ABXDYFGH	✓	✗
ABCDEFG CDEFG	✓	✗
DEFG DEFG	✓	✗
ABCDEF CDEF	✗	✗

The values 2 and 10 have been chosen as initial thresholds based on the average lengths of sequence pairs. As one can see, by defining these thresholds the similar pairs match the NW rule. However, some others pairs such as (ABCDEF,

CDEF) does not. If we want sequence pairs to have a similarity score higher than half of the sequence length, the longer sequences length has to be used within the rule. Integrating this length also allows overcoming another drawback of NW similarity metric as NW scoring scheme counts correlation between similarity and dissimilarity but ignores the ratio of similarity/dissimilarity over sequence lengths. Thus, we enhanced the rule to make NW score value dependent of the longer sequence length as described in Table 2, with the corresponding coefficient equals to 1/4 then all pairs match:

Table 2. Rule matching and non-matching pairs in sequence alignments result after taking longer sequence length into account through its coefficient

	NW score > longer sequence length/4	NW score > longer sequence length/4 and SW score > 10
ABCDEFG BCDEFG	✓	✓
ABCDEFGH ABXDYFGH	✓	✗
ABCDEFG CDEFG	✓	✗
DEFG DEFG	✓	✗
ABCDEF CDEF	✓	✗

However, by applying SW rule as threshold for the expected consecution, many pairs in NWs result are non-matching with this rule. We analyze the non-matching pairs as following:

- ABCDEFGH/ABXDYFGH: Resulting SW score = 10 when aligning with ABCDEFGH or other sequences because of its inner dissimilarity comparing to the other ones. This web access sequence is not similar to the other one in pair because the consecution is not matching SW rule
- CDEFG/DEFG: Resulting SW score = 8 when aligning with the other sequence or itself because of one disadvantages of this approach: SW score set in rule affects the sequence lengths in result, because these lengths have to be equal or greater than the threshold. Nevertheless, this can be improved by setting the SW score in the rule dependent of the shorter sequence of the set. For example, in order to select pairs that shorter sequence are sub sequence of longer one, similarity length aligned by SW must equal to the shorter sequence length.

With the above given matching score of SW aligning is 2, we change the rule condition from "> 10" to "=shorter sequence length x 2". Corresponding result is in Table 3, which shows the final result of proposed combination of NW and SW:

Table 3. Rule matching and non-matching pairs in sequence alignment result after taking longer and shorter sequence length into account through their coefficients

	NW score > longer sequence length/4	NW score > longer sequence length/4 and SW score = shorter sequence length x 2
ABCDEFG BCDEFG	✓	✓
ABCDEFGH ABXDYFGH	✓	✗
ABCDEFG CDEFG	✓	✓
DEFG DEFG	✓	✓
ABCDEF CDEF	✓	✓

Another possible approach is binary Dynamic Time Warping (DTW). Back to sequence pair examples from Figure 1 to 5, the application of DTW results are close to NW. If the rule is, for example, DTW score ≤ 2, pairs in Figure 2 to 5 are similar and pair in Figure 1 is not. The combination of DTW and SW returns a similar result than NW and SW when sequences pair in Figure 5 eliminated from similarity set of pairs. In DTW, conditional value in rule can depend on sequence length too, since the sequences pair considered similar if the dissimilarity not greater than some threshold

For instance, (AAAA,A) is a case that could not be considered similar in web usage mining context because there might be a reason why a web visitor stayed longer on a page. Nevertheless, DTW does not align with gaps as NW; hence it treats sequence of identical symbols not as a kind of user accessing behavior but as duplication. Therefore, DTW scores is 0 for this example, no matter how long is the duplication in the longer sequence. This limitation makes NW more suitable than DTW in page visit sequence alignment.

Time and Space Complexity: According to Alexander Chan in [12], time and space complexity of NW and SW are the same, $O(mn)$, given by m and n are sequence lengths. In our proposed method, each sequence pair is aligned by both of algorithms, thus the total time and complexity processing each pair should be $O(mn)$.

3 Experimental result

The dataset used for the experiments was collected from a University campus website. This website has more than 20,000 visits monthly. A deployed service

has taken part in preparation phase [13, 14] of the clustering process. Written in Javascript and Java, these services allow us to extract information from University campus data like cookies and other associated information such as page visit order, activity time or duration of page visit. In addition, the output format is optional which is convenient to work with variety of mining tool if needed. The extracted information is then checked and validated before applying algorithms to mine them.

Building web access sequences is the next phase. As mentioned earlier and in related works [3–7], sequence of visits plays an important role in user behaviors analyzing. In order to improve the performance of sequence alignment, URLs have to be shortened optimally by the presentation of symbols set like numbers. Similar to [5–7], session contains ordered URLs like, for example:

1 = `http://www.campus-fonderie.uha.fr/fr/droit/`
2 = `http://www.campus-fonderie.uha.fr/fr/economie-et-societe/`
3 = `http://www.campus-fonderie.uha.fr/fr/management/`
4 = `http://www.campus-fonderie.uha.fr/fr/management-interculturel/`

will be represented by symbol set $\{1, 2, 3, 4\}$, and turn to be page visit sequence like S = 1_2_3_4. In this sequence, page access order is respected and each symbol represents only a unique page. Pairwise alignments are made through all pairs of page visit sequence to score their similarity. As proposed in [3], similarity matrix of web sessions is then computed from this pairwise alignment results.

In the first experiment, we show results on 32 sample sessions that have been selected according to their length, duplication and order of visits as representative of the whole dataset. The goal of this experiment is to highlight specific features of the method. We focus this experiment on three rules: "NW score > longer sequence length/4" (NW), "SW score = shorter sequence length x 2" (named SW), and the combition "NW&SW" (the rule in the last column of Table 3). We applied independently the three rules on the similarty matrix obtained by comparing the 32 sequences. We then computed single linkage clustering using the three matrices using R[1]. The clustering results are displayed using dendrograms on Figure 6 for NW, 7 for SW and 8 for NW&SW

As we can see on Figure 6, 7 and 8, there are respectively 26, 32 and 23 sessions after the applications of the rules. Applying NW rule results (Figure 6) leads to more similar sessions with higher global similar, but sequences like 10_8_1_9_2_4 or 1_2_3_4_5 are not locally similar to others by SW rule. In contrary, applying only SW rule (Figure 7) leads to the existence of sequences such as 10_1_12_13_4_9_14 9_3_4 11_11_11,10_8_15_10, 10_8_1_9_2_4, etc. that are not globally similar to others by NW rule.

A noticeable sequence 9_3_4 exists in single rule cases because it is matching either but not both. Finally, the combination of NW and SW (Figure 8) rules extracts less but satisfied sequences of global and local similarity.

[1] `https://stat.ethz.ch/R-manual/R-patched/library/stats/html/dendrogram.html`

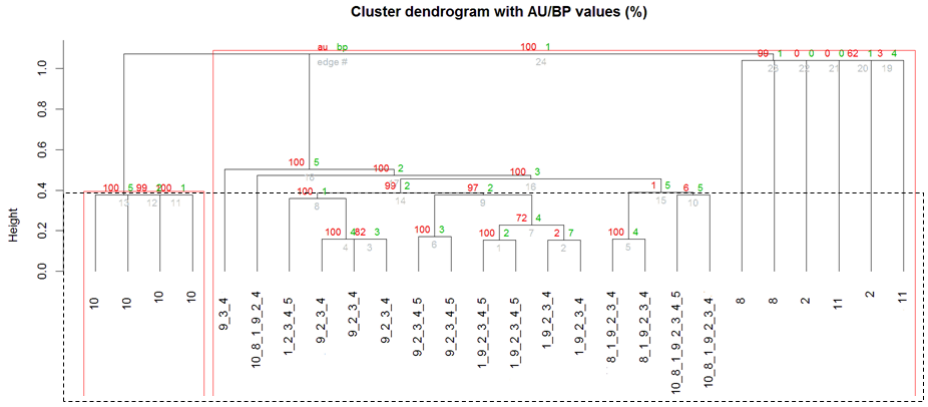

Fig. 6. Dendrogram of NW score > longer sequence length/4 (NW)

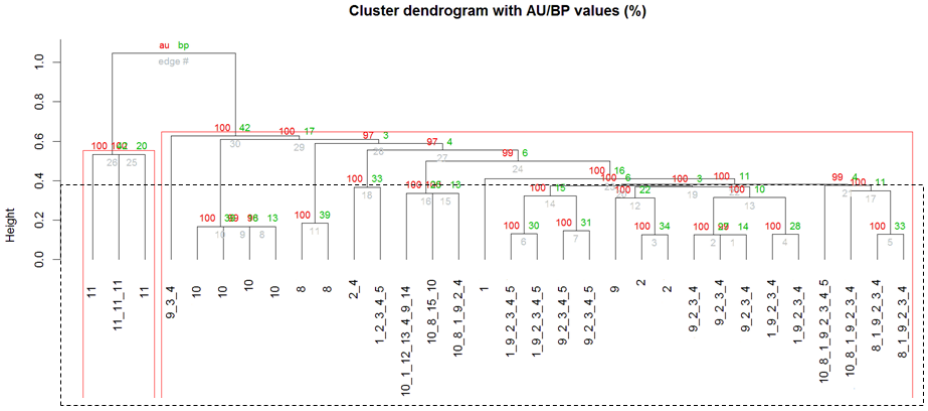

Fig. 7. R2: Dendrogram of SW score = shorter sequence length x 2 (SW)

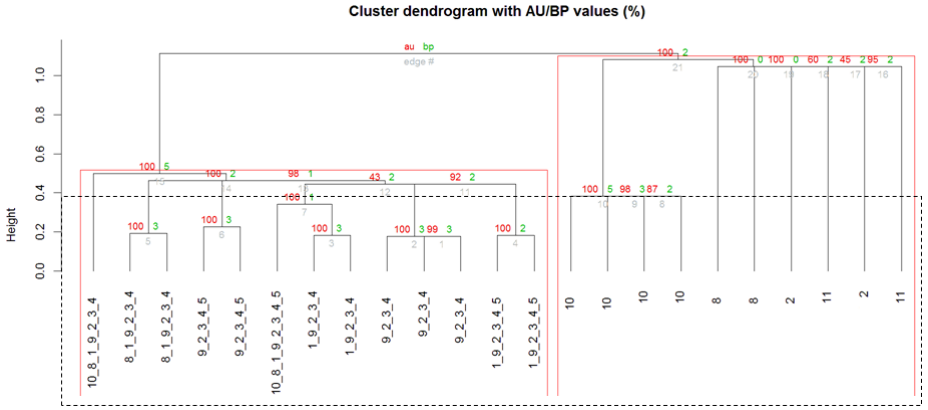

Fig. 8. Dendrogram of NW score > longer sequence length/4 and SW score = shorter sequence length x 2 (NW&SW)

Consequently, the first top-down pair of Figure 8 is different from the others on global and local similarity. Considering clusters at a specific height, for instance 0.4, by looking inside dashed frame, some following features can be seen:

- The number of clusters produced using NW, SW and NW&SW rules are respectively 11, 10 and 13 ;
- The NW&SW similarity inside the clusters is the best, most of the clusters are 100% similar inside, except one. Meanwhile, there are two and three of such clusters in NW and SW. The rate of clusters with 100% inside similarity over number of cluster are respectively 92%, 81% and 70% in NW, SW and NW&SW ;
- Clusters with dissimilarity inside of NW&SW are smaller than NW, and such clusters in NW are generally smaller than them in SW, considering the cluster size by number of sequence ;
- Clusters with dissimilarity inside of NW&SW are more similar than NW, and such clusters in NW are generally more similar than them in SW, considering the rate of similar sequences over number of sequences ;
- The necessary hierarchical level number by NW, SW and NW&SW rules are respectively 24, 30 and 21.

Experiments performed in the sample of 32 sessions show that the combination of NW and SW rules eliminates dissimilar sequence pairs from the similarity matrix compared to single NW and SW rules. As described previously, the combination of NW and SW rules generates more clusters from less sequence, more clusters 100% similar inside, even smaller and better similarity inside of dissimilar clusters, requires less hierarchical level than single NW and SW rules.

We also performed a similar experiment on another dataset containing 1128 page visit records of 282 sessions. We used the implementation proposed in [2], agglomerative hierarchical clustering algorithm in [3], and implement an algorithm to get clusters by optional level on tree. We also include the "no rule" when working on similarity matrix, to compare the number of clusters and execution times among rules application as in Table 4, 5 and 6, that present result after one of about fifty performing times of this experiment. Thus, Table 4 shows the number of clusters at some specific hierarchical levels by applying no rule and rule of "NW score > longer sequence length/4". Correspondingly, Table 5 is about cluster number by applying rules of "SW score = shorter sequence length x 2" and "NW score > longer sequence length/4 and SW score = shorter sequence length x 2" at same levels. Finally, in Table 6 we present execution times after running on corresponding rules. This result presents all mentioned advantages of NW and SW rules combination, as in the previous experiment with 32 sample sessions. Additionally, the execution times are also in the same order of "No rule" > "SW rule" > "NW+SW rule" > "NW rule" like Table 6. Therefore, the rules combination of NW and SW is better than others in clustering exactness, with some difference in execution time.

[2] `https://code.google.com/p/himmele/source/browse/trunk/Bioinformatics/`
[3] `https://github.com/lbehnke/hierarchical-clustering-java/tree/master/src`

Table 4. Number of clusters on hierarchical tree at some specific levels, by no rule and NW rule.

	No rule	NW score > longer sequence length/4
Hierarchical level 5	8	17
Hierarchical level 7	11	34
Hierarchical level 9	20	49
Hierarchical level 11	33	63

Table 5. Number of clusters on hierarchical tree at some specific levels, by SW rule and rule combination of NW and SW

	SW score = shorter sequence length x 2	NW score > longer sequence length/4 and SW score = shorter sequence length x 2
Hierarchical level 5	12	18
Hierarchical level 7	21	37
Hierarchical level 9	29	49
Hierarchical level 11	33	67

Table 6. Clustering execution time by no rule, NW rule, SW rule and rule combination of NW and SW

	Execution Time (in second)
No rule	692.7
NW score > longer sequence length/4	558.9
SW score = shorter sequence length x 2	578.0
NW score > longer sequence length/4 and SW score = shorter sequence length x 2	561.3

4 Related work

Our approach focuses on sequence alignment and ignores URL structural similarity suggested in some previous approaches [3, 4]. The reason is, without content mining, the use of website tree structure to represent the similarity of user interest may practically encounter some shortcomings such as:

- URLs are not always in any fixed structure form. Nowadays, they tend to be shorten by some unstructured presentation string
- The similarity of URL structure does not completely reflect the common interest of visitors. Two pages like `math.html` and `art.html` probably explored by two separated visitors groups and maybe shown in different categories on website, though they got the same prefix

In another approach [5], SW might be more global by counting the longer sequence length in pairwise alignment but our paper focuses on the combination of primitives. Proposed sequence alignment methods in [6, 7], using a hybrid metric by incorporation global into local alignment of 2 sequences. According to the formula, the more different in sequences length, the more local alignment should be taken into account, and vice versa. Because it turns to be global for the shorter sequence and local for the longer sequence, this metric is meaningful in some contexts. Nevertheless, it is not really in ours, because a local alignment scoring scheme like SW not counting the rest different length of sequences, although the longer these parts length is, the more important it will be in similarity metric when sequence lengths are significantly different.

Using similar pairwise alignment implemented by dynamic programming, DTW optimally minimizes the cost function [9, 10] ie. distance between pair of sequences whereas NW optimally maximizes similarity score. As a result, DTW measures the dissimilarity between sequences. In our context of web usage mining, two sessions with less dissimilarity in page visits are more similar, then DTW can be taken into account in considering proposed methods of sequence alignment. As our DTW analysis result above, DTW is appropriate for time series stretching or compressing but not for strings like our approach.

5 Conclusion

Sequence alignment techniques have been used widely in DNA sequences comparison, and have also been applied to segmentation of Web sessions. However, these techniques were not originally dedicated to web usage clustering, and there is room for optimization in order to adapt these alignments techniques to the specificities of real-time Web marketing, which is our field of application.

We have made the choice of a simple threshold-driven combination of the well-known Needleman-Wunsch and Smith-Waterman global and local alignment techniques. Values of these thresholds can be considered parameters of a given Web site, and we follow currently some simple heuristics to define them.

Our experiences show that our pairwise distance metric, based on the successive alignment of NW and SW in sequence pair, is a simple and realistic way to combine global and local approaches.

With the raise of mobile devices and tablets, there is now a significant difference in terms of low-level events that can observed between those devices and traditional computers (with a mouse). We need to better understand how the granularity of the events included in the sequences affects these thresholds. Therefore, future work is needed to fine-tune our heuristics for setting thresholds.

References

1. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology **48**(3) (1970) 443–453
2. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. Journal of molecular biology **147**(1) (1981) 195–197
3. Wang, W., Zaïane, O.R.: Clustering web sessions by sequence alignment. In: Database and Expert Systems Applications, 2002. Proceedings. 13th International Workshop on, IEEE (2002) 394–398
4. Li, C., Lu, Y.: Similarity measurement of web sessions based on sequence alignment. Wuhan University Journal of Natural Sciences **12**(5) (2007) 814–818
5. Poornalatha, G., Raghavendra, P.: Alignment based similarity distance measure for better web sessions clustering. Procedia Computer Science **5** (2011) 450–457
6. Chordia, B.S., Adhiya, K.P.: Grouping web access sequences using sequence alignment method. Indian Journal of Computer Science and Engineering (IJCSE) **2**(3) (2011) 308–314
7. Dimopoulos, C., Makris, C., Panagis, Y., Theodoridis, E., Tsakalidis, A.: A web page usage prediction scheme using sequence indexing and clustering techniques. Data & Knowledge Engineering **69**(4) (2010) 371–382
8. Petitjean, F., Forestier, G., Webb, G., Nicholson, A., Chen, Y., Keogh, E.: Dynamic time warping averaging of time series allows faster and more accurate classification. In: IEEE International Conference on Data Mining. (2014)
9. Meesrikamolkul, W., Niennattrakul, V., Ratanamahatana, C.A.: Shape-based clustering for time series data. In: Advances in knowledge discovery and data mining. Springer (2012) 530–541
10. Nakamura, A., Kudo, M.: Packing alignment: alignment for sequences of various length events. In: Advances in Knowledge Discovery and Data Mining. Springer (2011) 234–245
11. Marascu, A., Khan, S.A., Palpanas, T.: Scalable similarity matching in streaming time series. In: Advances in Knowledge Discovery and Data Mining. Springer (2012) 218–230
12. Chan, A.: An analysis of pairwise sequence alignment algorithm complexities: Needleman-wunsch, smith-waterman, fasta, blast and gapped blast. (2013)
13. Cooley, R., Mobasher, B., Srivastava, J.: Grouping web page references into transactions for mining world wide web browsing patterns. In: Knowledge and Data Engineering Exchange Workshop, 1997. Proceedings, IEEE (1997) 2–9
14. Cooley, R., Mobasher, B., Srivastava, J.: Data preparation for mining world wide web browsing patterns. Knowledge and information systems **1**(1) (1999) 5–32