Using Glocal Event Alignment for Comparing Sequences of Significantly Different Lengths

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

Using Glocal Event Alignment for Comparing Sequences of Significantly Different Lengths

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

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

Abstract. This work takes place in the context of conversion rate optimization by enhancing the user experience during navigation on e-commerce web sites. The requirement is to be able to segment visitors into meaningful clusters, which can then be targeted with specific call-to-actions, in order to increase the web site turnover. This paper presents an original approach, which equally combines global- and local-alignment techniques (Needleman-Wunsch and Smith-Waterman) in order to automatically segment visitors according to the sequence of visited pages. Experimental results on synthetic datasets show that our approach out-performs other typically used alignment metrics, such as hybrid approaches or Dynamic Time Warping.

Keywords: web mining, sequential pattern mining, clustering

1 Introduction

Conversion rate optimization is considered as one of the most promising approaches for improving the turnover of e-commerce web sites. A lot of researches have already focused on understanding web browsing event-patterns, in order to improve the online content delivery. Clustering visitors into meaningful segments, associated to targeted call-to-actions and related item recommendation, is one of the techniques typically used for cross- and up-selling. As clustering aims at organizing similar items into the same group with no prior knowledge of item class, it is seen as an approach of unsupervised learning. In web usage mining context, the similarity of page visits and their order in a session is one of the relevant information to cluster. For example, cluster analysis helps to reach people who are interested in some specific kind of goods or services so that the owner can recommend to such groups other related things, or offer them some discounts. The clustering result can also be applied to advertising placement organization on web sites, based on page visiting frequency in each cluster.

In this paper, we present our work for computing the similarity between event-sequences of significantly different lengths. Our proposal is based on a new

This is a pre-print of an article published at MLDM 2016. The final authenticated version is available online at: http://link.springer.com/chapter/10.1007%2F978-3-319-41920-6_5

way of equally combining global- and local-alignment techniques (*i.e.* Needleman-Wunsch [21] and Smith-Waterman [29]). The originality of our measure is to take into account the length of longest sequence in the pair of compared sequences.

Thus, regardless of the difference in sequence lengths, the result provided by our metric is accurate and can be used to perform clustering. Experimental results show that our approach outperforms other typically used similarity measures, such as hybrid approaches or Dynamic Time Warping (DTW), in the context of event-sequences of different lengths. This paper is divided into five sections with the following structure: Section 2 explains the proposed method. Section 3 describes experimental results. The discussion of these results is in Section 4. Section 5 presents related work. Finally, Section 6 concludes the paper and gives some future research directions.

2 Proposed method

Before discussing our sequence alignment approach, we introduce a few basic concepts: Given a finite set Σ whose elements are characters, called alphabet, any possible string of length $k > 0$ over Σ is a k-tuple built by characters from Σ. For example, if $\Sigma = \{A, B, C\}$, a set $S = \{s_1, s_2, \ldots, s_n\}$ of n finite strings over Σ can consist of $s_1 = AB, s_2 = ABC, \ldots, s_n = ACB$. In our model, each web session contains a series of page visits is assumed to be a sequence (*i.e.,* visits which are ordered). Hence, each sequence s_i is composed as a string from Σ, representing a session. A set of navigation sequences S, as mentioned, contains sessions from multiple visitors. To group sessions based on visit order of visitors, our method works as follows: S is processed to create clusters containing comparable sequences that are dissimilar to sequences in other clusters. For this purpose, our alignment-based similarity measure is proposed. An alignment over a set of sessions $S = \{s_1, s_2, \ldots, s_n\}$ can be described as another set $S_a = \{s_{1a}, s_{2a}, \ldots, s_{na}\}$ of equal length sessions which built by adding necessary gap "-"to s_i, for $1 \leq i \leq n$ [7]. Next, elements at the same *index* of session strings are compared and scored by a scoring scheme, so-called similarity definition.

Sequence similarity definition of specific context (or application-dependent) is essential to perform a relevant similarity evaluation. For instance, the correspondence of DNA sequences [26] is not identical to time-series [18], and both of them are different to web session similarity. Therefore, session similarity measure has to be adapted to web usage situation. At first, it has to deal with the variety of session lengths and thus traditional vector distances like Euclidean, Manhattan or even Hamming cannot be applied. Such distances require equal-length sequences like $s_1 = ABCD$, $s_2 = ABCD$. Secondly, as sessions are expected to be differentiated by page visit orders, the appropriate metric has to take into account this order. Thus, metrics such as Levenshtein and VLVD [25] are inappropriate as they consider the visit of page$A$ before B and vice versa to be the same. As a result, $s_1 = ABCD$ and $s_2 = BDCA$ are identical. These statistical approaches count the occurrence of element in each sequence to measure their similarity, regardless of element order. Additionally, the continuity of common pages between two browsing behaviors is a significant factor to evaluate their correspondence. Therefore, LCS or SAM[11] should not be used as it consider sessions started by page A and ended by B, that are common pages, like $s_1 = AB$ and $s_2 = ACDFB$ to be the same, regardless of how many unique pages between them. As the matter of fact, there is a meaningful difference in web visitor interest when one hits C,D and F between A and B and the other hits no pages between those two but they are not counted in this kind of metrics. In summary, an applicable approach in web usage mining context should be able to (1) process sequences with variable length, (2) take the order and succession of common pages into consideration. Such measures are suitable to compute the similarity between two sets of visits.

The Needleman-Wunsch (NW) method is a dynamic programming algorithm for sequence alignment which was developed by Needleman and Wunsch in 1970. Dynamic programming makes it possible to find the optimal alignment of sequences, is easy to implement and popular in computer science. When aligning elements of a sequence, matching and mismatching scoring scheme are given. A corresponding score matrix is then established to find the highest score of all possible alignments. In NW, this alignment is carried out from beginning to end of each sequence, it is called a *global alignment*. Global alignment is appropriate to work with sequences of similar length to find their best alignment. However, sequences may inherently not have the same length but might contains similar subsequences. Thus, a *local alignment* is relevant to detect them. To address this issue, the Smith-Waterman (SW) method, introduced by Smith and Waterman in 1981, performs the alignment by taking high comparable regions within sequences into account, regardless of the dissimilar parts and even the difference of sequence lengths. These two dynamic programming algorithms are commonly adopted for aligning protein or nucleotide sequences [14,19].

As NW and SW alignment methods are somehow opposite, each one has their own advantages and drawbacks. NW finds the optimal similarity of the entire sequence, while SW detects regions of likeliness between two sequences. As a result, a combination of both methods is better than using a single one, since the correspondence between sequences can be evaluated correctly (*i.e.* globally and locally). We pointed out the effectiveness of the NW and SW rules combination compared to DTW [24] and hybrid metric [8,6] in our previous work [16]. Following this finding, we propose a new similarity metric called *combination metric*. This metric is based on NW, SW and the size of the longest sequence:

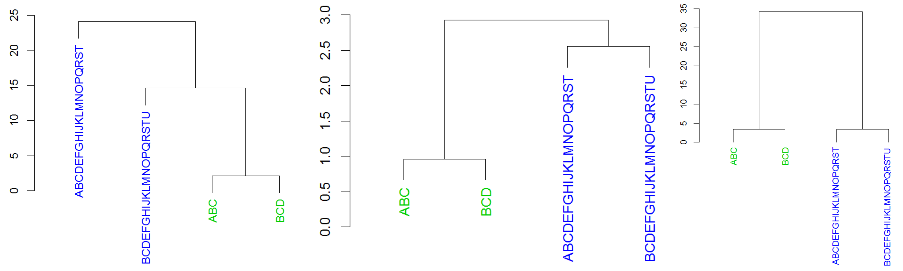
$$S(s_i, s_j) = \left[\frac{NW(s_i, s_j)}{l}\right] + \left[\frac{SW(s_i, s_j)}{(2 * l)}\right] \tag{1}$$

with $NW(s_i, s_j)$ and $SW(s_i, s_j)$ respectively NW and SW scores between the two sequences s_i and s_j, l the length of longest sequence in the pair (*i.e.* $max(|s_i|, |s_j|)$), the NW scoring scheme of +1 for matching and -1 for non-matching pair of items in sequences, the SW scoring scheme of +2 for matching and -1 for non-matching inside matching, ignore non-matching outside.

Another way to combine NW and SW was proposed previously in *hybrid metric* scores similarity [8,6] between two sequences s_i and s_j, is defined as:

$$S(s_i, s_j) = (1 - p) * SW(s_i, s_j) + p * NW(s_i, s_j) \tag{2}$$

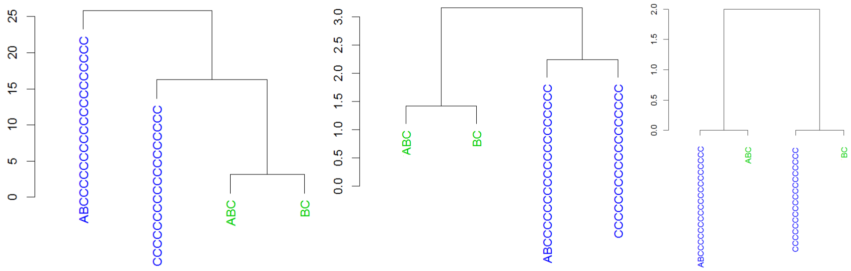
with the defined parameter $p = |s_i|/|s_j|$. From this definition (Eq. 2) , one can notice that hybrid metric does not equally take both NW and SW into account because of the difference in sequence lengths. As a consequence, the advantage of *combination metric* (Eq. 1) over the *hybrid metric* (Eq. 2) is that the similarity measure works better when sequence lengths are very different. In this case, hybrid metric only focuses on SW, while combination metric focuses on both NW and SW. Consequently, the more different in lengths the sequence pair are, the more the hybrid metric will focus on SW to measure their similarity. Therefore, the hybrid metric would consider two sequences of different classes to be in same class, if their difference in length and SW are important enough. On the other hand, two sequences of same class with comparable lengths may not be similar enough to be in the same cluster because their similarity score by hybrid metric is smaller than in the previous case. To illustrate this scenario, we created cluster dendrograms with a toy example. The Figure 1a shows the similarity evaluation of the hybrid metric in case of sequences with quite different length. As the two sequences of blue class are not similar enough to be merged in agglomerative hierarchical clustering, the resulting clustering is of poor quality. As illustrated in Figure 1b, this case does not happen using the combination metric as both NW and SW are considered equally, regardless of the length difference between the sequences. Consequently, two clusters green and blue of perfect accuracy are obtained when the dendrogram is cut.



(a) Hybrid metric. (b) Combination metric. (c) DTW

Fig. 1: Example of clustering of 4 sequences of 2 classes (blue and green) with quite different length for hybrid (a), combination (b) and DTW (c) metrics.

Dynamic Time Warping (DTW) is known to be effective to find good alignment between time-series. In the context of sequence pairs of quite different lengths and no duplicate as in Figure 1, DTW works mostly as good as combination method. As illustrated in Figure 1c, it makes blue and green sequences merged in agglomerative hierarchical clustering. However, DTW minimizes distance of one sequence to another by allowing flexible transformation so that time-series with similar shapes can be detected. This feature leads to a problem when identical consecutive elements in sequences are merged. Thus, the warping path of sequence pairs is vertical or horizontal. In other words, a single element from one sequence is aligned with many successive and duplicate elements in the other as they are all identical. This feature is a drawback in web usage mining as sessions containing duplicate web pages should not be "skipped" but mined for web visitor behavior. Our proposed metric considers them to be traversal pattern and takes this duplication into account while DTW regards them as only one page no matter how many visits are duplicated. Figure 2b and 2c respectively illustrate the similarity evaluation of the combination metric and DTW in case of sequences with duplicate elements. Furthermore, with sequence pairs which contain duplicate elements and quite different in length like in Figure 2, hybrid metric is less effective than the combination, as presented in Figure 2a.



(a) Hybrid metric. (b) Combination metric. (c) DTW

Fig. 2: Example of clustering of 4 sequences of 2 classes (blue and green) with duplicated elements for hybrid (a) and combination (b) and DTW (c) metrics.

3 Experimental results

3.1 Synthetic data

In order to evaluate the performance of our combination metric, clustering validation including internal and external measures are considered. As some appropriate classified data is not currently available for external validation, we used internal validation on generated synthetic datasets. Nevertheless, Liu et al. [13] analyzed both kind of validations and revealed the relevance of internal validation measure in many aspects over some other external validation measures. Rendon et al. [27] also concluded that they can get better precision by validity internal indexes than external ones, on their datasets and scenarios. To evaluate the performance of the proposed metric and competitors (*i.e. hybrid metric* and *DTW*), we generated 10 synthetic datasets randomly. Each of them contains more than 500 sequences of sessions (about 520 in average) which are grouped into three defined classes with following features:

- Class 1: About 170 sequences of lengths $[20 - 22]$, sharing a common subsequence, for instance: A**BC**DU3YU31DQ6Q4FO2JGHW, A**BC**DSI5OPHH9EDGLPFLST, A**BC**DAFF5UAK7GEX3XJIU. Generated sequences in other classes are related to this common sub sequence ;

- Class 2: About 170 sequences of lengths $[3 - 4]$, sharing a common subsequence (that is also sub-sequence of common sub-sequence in Class 1), for instance: Z**BC**, A**BC**5, **BC**0 ;

- Class 3: About 170 sequences of lengths $[18-20]$, mostly containing identical and consecutive symbols (that appear in the common sub-sequence in Class 1), for instance: DDDDDDDDB**BC**CCCCCCC, DDDDDDDD**C**CAAAAAAA, BBBBBBB**B**BADBBBBBBBBB ;

Note that all the datasets used in the experiments are available to download here[1]. As shown in Section 2, sequences with common subsequence but different lengths such as Class 1 and 2 are likely to be misclassified by hybrid metric. Yet sequences with same consecutive symbols in Class 3 are likely to be confused with sequences of Class 2 using DTW. These classes are assumed to be representative of behaviors that can be witnessed when analyzing real web sessions.

Using these sequence datasets, similarity matrices for each metric were computed. In order to implement agglomerative hierarchical clustering [9], these matrices are then used as input with three well known hierarchical methods: *single-linkage* (sing.), *complete-linkage* (compl.) and *average-linkage* (avg.). This variety of hierarchical methods contributes to the effectiveness of the evaluation. Table 1 shows the means (μ) and standard deviations (σ) of clustering result precision for the three methods with the three hierarchical methods over the 10 datasets. Note that as hierarchical clustering is deterministic, running the experiments multiple times is not required. Thus, the means and standard deviations correspond to the execution on the 10 different datasets.

Table 1: Results for the three methods on the 10 datasets.

	Original datasets								
	Hybrid			DTW			Combination		
	sing.	compl.	avg.	sing.	compl.	avg.	sing.	compl.	avg.
μ	**100%**	**58.5%**	**100%**	**90.5%**	**85.5%**	**89.7%**	**100%**	**98.2%**	**100%**
σ	±0%	±8.5%	±0%	±9.7%	±1.4%	±9.8%	±0%	±5.7%	±-0%

The correlation of experimental results in Table 1 illustrated by dendrograms in Figure 3, 4 and 5 on sample set of the sequences (for sake of clarity) with leave values as defined classes. By cutting dendrograms at the desired level, clusters are separated into frames and their quantity matching the number of defined classes. As shown in Figure 3, there is one cluster with unexpected accuracy containing

[1] https://www.dropbox.com/sh/b6wxv5opn1u3n6n/AAB8ObwvqBPbDsnXvB9xZ_yca

sequences from both Class 1 and 3. The number of inaccurate clusters in Figure 4 is two as they correspond to Class 2, 3 and Class 1, 3 sequences. However, clusters in Figure 5 achieve a perfect accuracy. These figures also illustrate that combination metric works very well compared to hybrid and DTW.

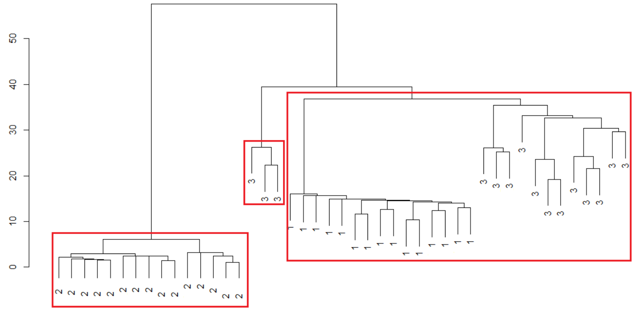


Fig. 3: Hierarchical clustering using hybrid metric on original dataset.

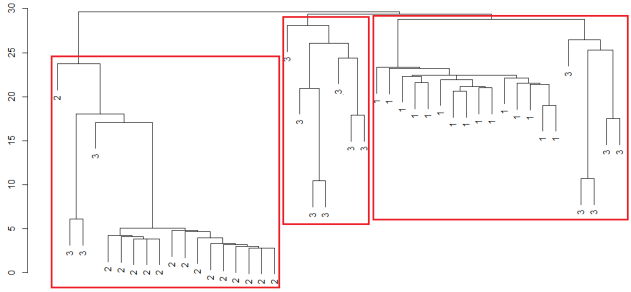


Fig. 4: Hierarchical clustering using DTW metric on original dataset.

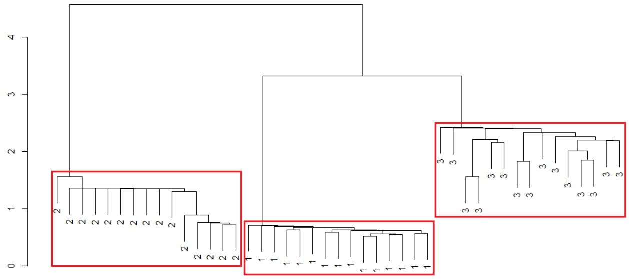


Fig. 5: Hierarchical clustering using combination metric on original dataset.

Following, we present additional results performed to consider two popular aspects of data in web usage mining: the noise and unbalanced density of classes (*i.e.* classes with important difference in number of elements).

Noise: About 15 sequences of lengths $[3, 24]$ were randomly generated from alphabet and numbers, for instance: APE8V98MDTIH77I, H96YXT7N, M9AKKAA, etc. were added to the original datasets with 3 classes. The accuracy of clustering results on these datasets is presented in Table 2.

Table 2: Results for the methods on the 10 datasets with noise.

	Datasets with noise								
	Hybrid			DTW			Combination		
	sing.	compl.	avg.	sing.	compl.	avg.	sing.	compl.	avg.
μ	**90.4%**	**84.3%**	**90%**	**65.8%**	**73.1%**	**86.7%**	**100%**	**89.2%**	**100%**
σ	±11.6%	±3%	±7.2%	±1%	±9%	±11%	±0%	±6%	±-0%

Similarly to the previous results, dendrograms on sample of sequences with leave values as defined classes are presented on Figure 6, 7 and 8. These figures illustrate the correlation of experimental results presented in Table 2. Dendrograms are cut at the desired level to make cluster quantity match the defined number of classes and separate them into red frames.

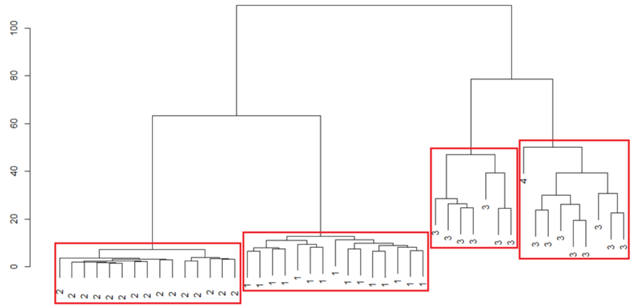


Fig. 6: Hierarchical clustering using hybrid metric on dataset with noise.

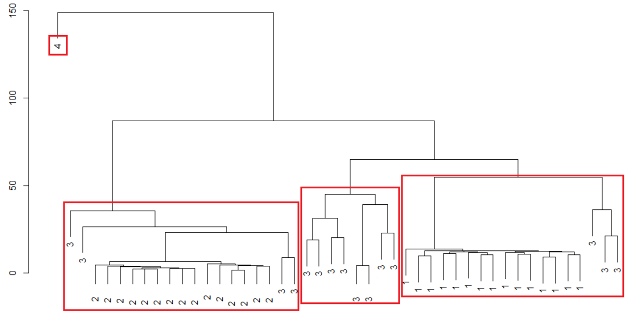


Fig. 7: Hierarchical clustering using DTW metric on dataset with noise.

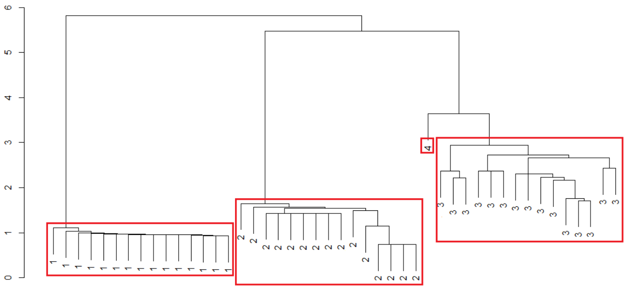


Fig. 8: Hierarchical clustering using combination metric on dataset with noise.

Unbalanced density: The number of sequences of Class 1, 2 and 3 are respectively around 320, 170 and 10 in the first three datasets. In the next four datasets, number of sequences in Class 1, 2 and 3 are respectively around 170, 320 and 10. Lastly, in the remaining three datasets, sequence numbers are around 10, 170 and 320 for Class 1, 2 and 3. As the number of users having the same usage can be very different according to specific behaviors, this kind of datasets are assumed to be representative of the data available in web usage mining. The results of the experiments using these unbalanced datasets are presented in Table 3.

Table 3: Results for the methods on the 10 datasets with unbalanced classes.

	Unbalanced density datasets								
	Hybrid			DTW			Combination		
	sing.	compl.	avg.	sing.	compl.	avg.	sing.	compl.	avg.
μ	**81.6%**	**73.2%**	**84.8%**	**90.9%**	**91.1%**	**91.2%**	**100%**	**93.7%**	**91.1%**
σ	±19%	±18.9%	±24.5%	±16.5%	±12.7%	±12.8%	±0%	±10%	±-0%

As mentioned above, experimental results in Table 3 are illustrated by sample data dendrograms in Figure 9, 10 and 11, with leave values as defined classes. Similarly, red frames separate elements into class defined number of clusters by cutting the tree at desired levels. Similar to the previous dendrograms, it presents the best accuracy obtained using combination metric compared to the others.

3.2 Real data

The dataset used for our external validation was collected from a commercial website. The dataset was provided by the Beampulse company which commercializes a product written in Javascript and Java, which collects information about web visitors behaviours such as page visit order, activity time or duration of page visit. As shown in dendrograms in Figure 12,13 and 14 that is a sample extracted from experimental result on 1500 individual sessions, each contains numbered page(s), the advantages of our metric is highlighted compared to the

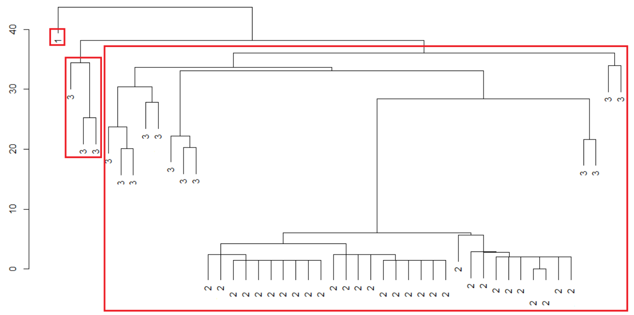


Fig. 9: Hierarchical clustering using hybrid metric on unbalanced dataset.

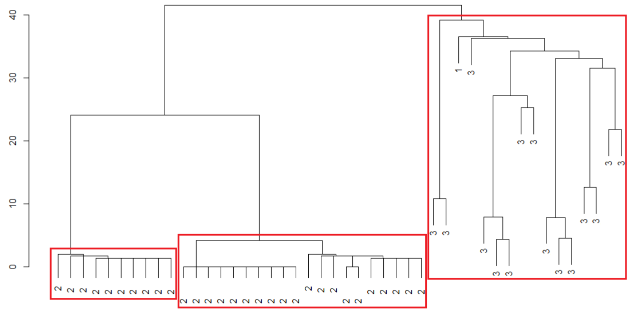


Fig. 10: Hierarchical clustering using DTW metric on unbalanced dataset.

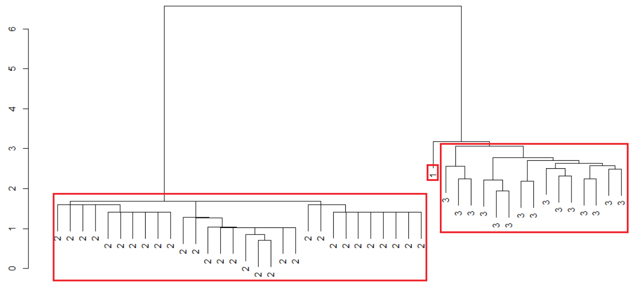


Fig. 11: Hierarchical clustering using combination metric on unbalanced dataset.

other methods. Using our metric, sessions with similar pages, similar page order and similar length are likely to be grouped in the same cluster. However, such features are not obtained using hybrid and DTW metrics.

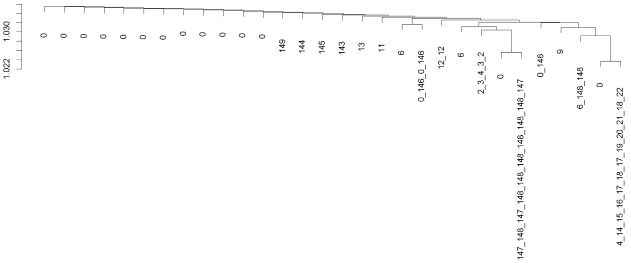


Fig. 12: Hierarchical clustering using DTW metric on real dataset

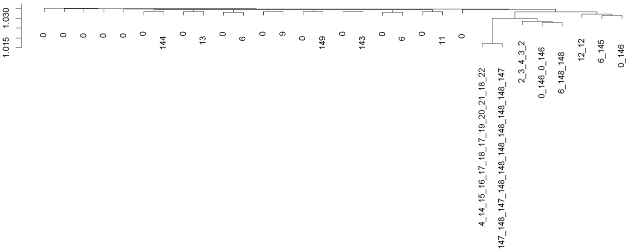


Fig. 13: Hierarchical clustering using hybrid metric on real dataset

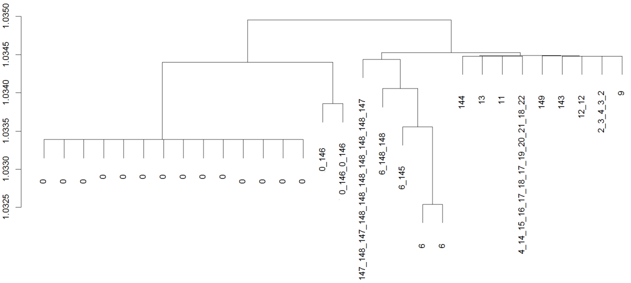


Fig. 14: Hierarchical clustering using combination metric on real dataset

4 Discussion

As shown in Table 1, throughout the 10 normal datasets (*i.e.,* with neither noise nor unbalanced clusters density), similarity matrix produced by DTW outputs the lowest precision clustering using single- and average-linkage. However, DTW precision is higher than hybrid metric by complete linkage. Hybrid metric is as good as combination metric using single and average-linkage but is significantly worse using complete-linkage where combination metric is mostly stable.

As noise may impact the clustering algorithm performance, it is good to obtain datasets without noise before clustering. However, clustering algorithms should have the ability to deal with noise because of the difficulty to avoid its presence, especially in big datasets. Our metric maintains a perfect precision using average and single-linkage, and also provides the highest precision using complete-linkage in Table 2, where 10 datasets including noise are used as input. Meanwhile, hybrid handles noise more accurately than DTW in this context.

Similarly, clustering methods are challenged by various density datasets because of its importance in the clustering process. On unbalanced datasets, the good clustering precision obtained using our metric remain almost the same using single and average hierarchical methods. In contrast to DTW, hybrid metric is highly influenced by unbalanced density. Again, our metric reached the best accuracy using single-linkage and always works better than the other two methods using complete and average linkage (see Table 3).

5 Related work

The goal of web usage mining is to identify hidden pattern from visitor browsing data. This involves clustering of different visits having similar navigational patterns. One of the most popular approaches to discover these clusters is session classification. Among various classification forms, many previous studies have focused on sequence alignment algorithms to evaluate the similarity of sessions. Mandal et al. [17] presented the calculation of distance between two sessions using Cosine measure but it requires sessions with exactly the same length, which is more suitable to vectors than web accesses. Furthermore, these approaches also ignore regions of local similarity of session pages. In [5], Chitraa et al. intended to find usage patterns using k-means algorithm application, yet clustering is to discover hidden pattern without specific input parameters.Similarly, defining k by the granularity that clusters should be in order to group web users proposed by Jianfeng et al. in [28] is inappropriate to browsing sessions.

Pairwise alignment is commonly used to compare sequences optimally. There exist sequence alignment algorithms, both global and local adopted through pairwise alignment such as NW [12] and SW [31]. These two algorithms take into account the similarity between sequences in different alignment [30] and have their own strengths and drawbacks [10]. Lu et al. [15] studied how to generate significant usage patterns using NW, however it ignores consecution that is essential to evaluate similarity of web session pairs. In contrary, a local alignment algorithm such as SW can only detect partial similarities [2].

Consequently, there have been previous works on integrating one into the other to take advantage of the combination. For example, Brudno et al. [4] proposed a system to align genomes with biological features glocally. Chordia et al. [6] described a hybrid metric, concerning a consolidation of global and local sequence similarity scoring. Correspondingly, the same metric was developed by

Dimopoulos et al. [8] to measure the similarity of two sequences. This formula computes the distance between sequences by taking global and local alignment and their weights into consideration. These weights are in inverse proportion to each other, depending on how different sequences are in length. Specifically, local alignment weight would be greater if sequence lengths are different. As local alignment scoring does not take the difference in sequence lengths into account, this computation may work in some specific situations but not in web accesses similarity because that difference reveals the dissimilarity in visitor browsing behavior. Similarly, Algiriyage et al. [1] used Levenshtein distance as a similarity metric for web session pairs but it does not identify the succession of common visits. On the other hand, with regard to sequential characteristic of web session, there has been approach such as [3] compares homogeneity of sequence pair by frequency of common items occurrence but does not take into account their order, that is key feature to differentiate sessions. DTW, that is a popular algorithm in comparing two sequences of events [20] or data points [23], is also used in symbolic sequence comparison. However, this kind of approach is not effective in web usage mining since DTW ignores duplicated elements. Consequently, it is not able to evaluate the dissimilarity in browsing behavior comparing a specific web page loaded many times with the same page loaded only one time.

Note that the lack of available benchmarks in the domain of web usage mining makes the comparison of the different existing methods difficult. In order to address this issue, we released with this paper all datasets[2] (*i.e.* original, with noise and with unbalanced density) that were used for the experiments. We hope that these datasets will be used in future research to compare new contributions.

6 Conclusion

In this paper, we have presented our contribution to event-sequence comparison, with a specific focus on sequences of significantly different lengths. This new way of combining global- and local-alignment techniques is based on the equal combination of both approaches. We experimentally evaluated this approach in context of clustering web site visitors, by analyzing the browsing patterns. Under those settings, it was observed that our sequence similarity metric outperformed other related techniques. In the close future, we plan to introduce mutations in the current datasets to better stress the combination technique in the presence of noise. We also want to compare the approach with other techniques such as PAM/k-medoids, ROCK or Ward. Furthermore, other distance measures such as Hamming or Levenshtein, which have been studied in a variety of sequence comparisons including spectra [22] should be considered in upcoming works.

Supplementary materials

All the datasets used in the experiments are available here: https://www.dropbox.com/sh/b6wxv5opn1u3n6n/AAB8ObwvqBPbDsnXvB9xZ_yca.

[2] https://www.dropbox.com/sh/bse1ifyu2gdywm4/AACRSbKysPVGudinjTBg6Ocsa

References

1. Algiriyage, N., Jayasena, S., Dias, G.: Web user profiling using hierarchical clustering with improved similarity measure. In: Moratuwa Engineering Research Conference (MERCon), 2015. pp. 295–300. IEEE (2015)
2. Aruk, T., Ustek, D., Kursun, O.: A comparative analysis of smith-waterman based partial alignment. In: Computers and Communications (ISCC), 2012 IEEE Symposium on. pp. 000250–000252. IEEE (2012)
3. Bouguessa, M.: A practical approach for clustering transaction data. In: Machine Learning and Data Mining in Pattern Recognition, pp. 265–279. Springer (2011)
4. Brudno, M., Malde, S., Poliakov, A., Do, C.B., Couronne, O., Dubchak, I., Batzoglou, S.: Glocal alignment: finding rearrangements during alignment. Bioinformatics 19(suppl 1), i54–i62 (2003)
5. Chitraa, V., Thanamni, A.S.: An enhanced clustering technique for web usage mining. In: International Journal of Engineering Research and Technology. vol. 1. ESRSA Publications (2012)
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), 308–314 (2011)
7. Della Vedova, G.: Multiple Sequence Alignment and Phylogenetic Reconstruction: Theory and Methods in Biological Data Analysis. Ph.D. thesis, Citeseer (2000)
8. 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), 371–382 (2010)
9. Duraiswamy, K., Mayil, V.V.: Similarity matrix based session clustering by sequence alignment using dynamic programming. Computer and Information Science 1(3), p66 (2008)
10. Giegerich, R., Wheeler, D.: Pairwise sequence alignment. BioComputing Hypertext Coursebook 2 (1996)
11. Hay, B., Wets, G., Vanhoof, K.: Clustering navigation patterns on a website using a sequence alignment method. Intelligent Techniques for Web Personalization: IJCAI pp. 1–6 (2001)
12. Likic, V.: The needleman-wunsch algorithm for sequence alignment. Lecture given at the 7th Melbourne Bioinformatics Course, Bi021 Molecular Science and Biotechnology Institute, University of Melbourne (2008)
13. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J.: Understanding of internal clustering validation measures. In: International Conference on Data Mining. pp. 911–916. IEEE (2010)
14. Liu, Y., Hong, Y., Lin, C.Y., Hung, C.L.: Accelerating smith-waterman alignment for protein database search using frequency distance filtration scheme based on cpu-gpu collaborative system. International journal of genomics 2015 (2015)
15. Lu, L., Dunham, M., Meng, Y.: Discovery of significant usage patterns from clusters of clickstream data. In: Proc. of WebKDD. pp. 21–24. Citeseer (2005)
16. Luu, V.T., Forestier, G., Fondement, F., Muller, P.A.: Web site audience segmentation using hybrid alignment techniques. In: Trends and Applications in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science, vol. 9441, pp. 29–40. Springer International Publishing (2015)
17. Mandal, O.P., Azad, H.K.: Web access prediction model using clustering and artificial neural network. In: International Journal of Engineering Research and Technology. vol. 3. ESRSA Publications (2014)

18. Meesrikamolkul, W., Niennattrakul, V., Ratanamahatana, C.A.: Shape-based clustering for time series data. In: Advances in Knowledge Discovery and Data Mining (PAKDD), pp. 530–541. Springer (2012)
19. Muhamad, F.N., Ahmad, R., Asi, S.M., Murad, M.: Reducing the search space and time complexity of needleman-wunsch algorithm (global alignment) and smith-waterman algorithm (local alignment) for dna sequence alignment. Jurnal Teknologi 77(20) (2015)
20. Nakamura, A., Kudo, M.: Packing alignment: alignment for sequences of various length events. In: Advances in Knowledge Discovery and Data Mining (PAKDD), pp. 234–245. Springer (2011)
21. 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), 443–453 (1970)
22. Perner, P.: A novel method for the interpretation of spectrometer signals based on delta-modulation and similarity determination. In: Advanced Information Networking and Applications (AINA), 2014 IEEE 28th International Conference on. pp. 1154–1160. IEEE (2014)
23. Petitjean, F., Forestier, G., Webb, G., Nicholson, A.E., Chen, Y., Keogh, E., et al.: Dynamic time warping averaging of time series allows faster and more accurate classification. In: International Conference on Data Mining. pp. 470–479. IEEE (2014)
24. Petitjean, F., Gançarski, P.: Summarizing a set of time series by averaging: From steiner sequence to compact multiple alignment. Theoretical Computer Science 414(1), 76–91 (2012)
25. Poornalatha, G., Raghavendra, P.S.: Web user session clustering using modified k-means algorithm. In: Advances in Computing and Communications, pp. 243–252. Springer (2011)
26. Qi, Z., Redding, S., Lee, J.Y., Gibb, B., Kwon, Y., Niu, H., Gaines, W.A., Sung, P., Greene, E.C.: Dna sequence alignment by microhomology sampling during homologous recombination. Cell 160(5), 856–869 (2015)
27. Rendón, E., Abundez, I., Arizmendi, A., Quiroz, E.: Internal versus external cluster validation indexes. International Journal of computers and communications 5(1), 27–34 (2011)
28. Si, J., Li, Q., Qian, T., Deng, X.: Discovering *k* web user groups with specific aspect interests. In: Machine Learning and Data Mining in Pattern Recognition, pp. 321–335. Springer (2012)
29. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. Journal of molecular biology 147(1), 195–197 (1981)
30. Yan, R., Xu, D., Yang, J., Walker, S., Zhang, Y.: A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. Scientific reports 3 (2013)
31. Zahid, S.K., Hasan, L., Khan, A.A., Ullah, S.: A novel structure of the smith-waterman algorithm for efficient sequence alignment. In: International Conference on Digital Information, Networking, and Wireless Communications (DINWC). pp. 6–9. IEEE (2015)