PROPOSED CLUSTERING MODEL BASED ON SEQUENTIAL RULES IN THE WEB MINING

[1]HADJ-TAYEB Karima
karima.hadjtayeb@univ-usto.dz
Université des sciences et de la technologie d'Oran
Mohamed Boudiaf USTO- Algérie

[2]BELBACHIR Hafida
h_belbach@yahoo.fr
Université des sciences et de la technologie d'Oran
Mohamed Boudiaf USTO- Algérie

***Abstract*—The important growth of the pages contained in the website as well as the number of users navigating requires research tools that allow studying the behaviour of users in the Web Usage Mining. Among these tools, clusters analysis is considered as the most important technique in this area. Based on this technique, several methods have been developed; the most popular is the partitioning method. However, its principle, as it appears to be unsuitable for web data, represents a sequential data stream where the similarity notion must be taken into account when calculating the distance between objects.**

This paper attempts to overcome the limitations of this method and proposes a new user clustering model. The proposed approach is based on the extraction of sequential patterns along with the generated sequential rules. The experimental realization has been carried out by implementing the proposed algorithm and the k-medoids partitioning algorithm. This study is carried out with the aim of comparing the performance relative for each of them through a set of measures that help evaluate the quality of the generated clusters.

Keywords— ***Web Usage Mining, Clustering Technique, Sequential patterns Technique, k-medoids algorithm; Sequential rules, Evaluation measures of clusters' quality.***

I. Introduction

To better understand the behaviour of web users and satisfy their needs, it is important to process and analyze this large data by applying the Data Mining techniques. Among these techniques, the cluster analysis is the most important technique used in Web data where time is primordial to follow visited pages. Among the existing methods of clustering, this article is interested in partitioning methods in the clustering of web users. It aims to provide solutions because this method presents some disadvantages when applied to web data. To this end, a new clustering model is proposed based on sequential patterns and rules.

As part of a comparative study of performance, this article compares the performance of a proposed approach with the performance of the popular partitioning k-medoid algorithm.
This study is carried out through a set of evaluation measures that allow evaluating the quality of the generated clusters in a minimum time.

II. Problem

Depending on how the clusters are formed and the types of processed data, five types of approaches have been proposed such as: hierarchical approach, partitioning approach, designed approach for high dimensional data, designed approach for categorical data, and Grid-based approach. **[1] [2]**
Among these approaches, the popular method is the partitioning method which subdivides objects into a number of classes by using an iterative optimization strategy whose principle is to generate the initial partition. The partitioning method improves progressively by allocating data from one class to another. Indeed, each object is assigned to the closest center by calculating the distance measure with the proximity between the centroid and the object. Thereafter, the new centers representing groups are recalculated. **[3] [4]**.
The algorithms for this method have the advantages to be: easy to understand and to implement, to be fast, to have a low requirement in memory size, and to be applicable to any type of data. The algorithms used in this method are: the k-means **[5]** algorithm, the k-medoids algorithm **[6]** and the EM algorithm **[7]**

Among these two methods, this article is based on the principle of the partitioning method to build a clustering model of web users. However, its principle as it appears to be unsuitable for web data for the following reasons:

The final partition depends on the initial partition due to the random selection of the k initial points that can generate a bad partitioning. These algorithms need to run multiple times with different initial states to obtain better results. In each initialization with a defined number of clusters, there are different solutions that can be away from the optimal solution. Then, it eventually becomes necessary to run these algorithms multiple times with different initializations and retain the best regrouping data. The use of this solution is limited because of its very high cost in terms of time calculation; the memory space and the number of steps as the best partition can be

obtained after several runs of the algorithm. These limitations make the application of these algorithms unsuitable on the large data bases.
- The calculated distance's measure between the cluster center and the affected object in partitioning approach seems inappropriate on the web data. Indeed, this distance is insufficient since it doesn't take into account the similarity between calculated objects, nor does it reflect the order of the items in the sequence. The cluster centroid, which is its representative in any way, reflects the behaviour browsers from the pages component and the cluster which is an aligned sequence.
- The partitioning algorithms can generate groups at risk of becoming empty during the regrouping process.

Among these partitioning algorithms, the k-medoids algorithm is considered as the most important algorithm in the partitioning approach. In the k-medoids algorithm, a class is represented by one of its points, which is called medoid. Such a representation has got two advantages: it adapts to any attribute type, besides, the medoid is chosen as a fraction of the predominant points in a class, so it does not become sensitive to outliers. Once the medoids are chosen, the classes are defined as the subsets of points close to the corresponding medoid and the objective is to minimize the distance between points labelled to be in a cluster and a point designated as the center of that cluster. It uses Manhattan distance to define the distance between data points.
The k-medoid algorithm follows the steps:
1. Initialization: it randomly selects k of the n data points as the medoids
2. Assignment step: it associates each data point to the closest medoid.
3. While the cost of the configuration decreases:
For each medoid m, for each non-medoid data point o:
- Swap m and o, recomputed the cost (sum of distances of points to their medoid)
- If the total cost of the configuration increased in the previous step, undo the swap.

The k-medoid algorithm repeats alternating steps 2 and 3 until there is no change in the assignments.
[6].

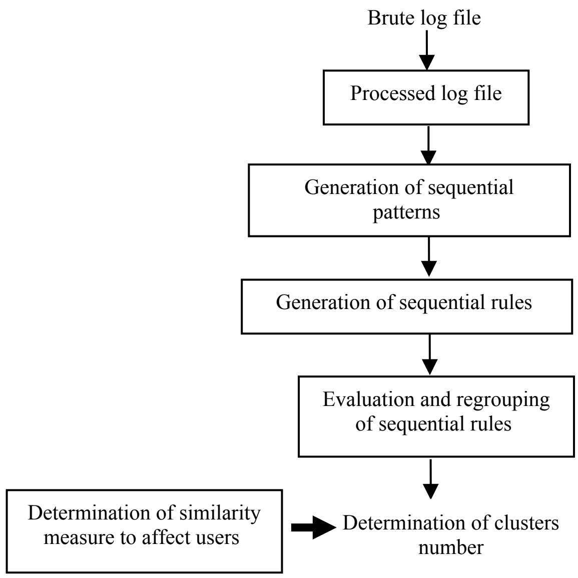
III. Related works

Based on the implementation of the k-medoids algorithm, several works have been done:

In [8], authors employ the EM algorithm in two clustering scenarios, for the construction of predictive Web Usage models. In the first scenario, user navigation paths are considered members of one or more clusters, and the EM algorithm is used to calculate the model parameters for each cluster. The probability of visiting a page is estimated by calculating its conditional probability for each cluster. The resulting mixture model is named Naive Bayes mixture model since it is based on the assumption that pages in a navigation path are independent given cluster. The second scenario uses Markov chains that represent the navigation paths of users clustered using the EM algorithm in order to predict subsequent pages.
In [9], authors propose a model of web navigators clustering based on Markov chain and on EM algorithm to discover semantic relationships between pages and users of each cluster. Each cluster in a Markov chain is composed of users' session that represents a behavior described by a single Markov chain. The assignment of user's session to a cluster is done by EM algorithm. To determine the clusters number, authors propose to use several probabilistic criteria (Bayesian criteria) where each cluster is generated from its own probability distribution.
In [10], authors illustrate a new method to extract navigation patterns by using an aligned sequence method. This method partitions a trace of navigations according to the order of required pages and administrates the problem of sequences clustering of different sizes on which is based the distance count. Experimental results are compared with a method based on the Euclidian distance that does not contain any information on the sequence.
In [11], authors propose an effective modification of k-means partitioning algorithm for the web sessions clustering. This clustering is effectuated by the distance function based on the vector of varying length of sessions. The order of visited pages is not taken into account in session's comparison. In this algorithm, the alignment number is considered for determining the distance between sessions.

IV. Proposed approach

The proposed approach tries to improve the performance of partitioning algorithms. It aims to eliminate random and iterative selection of the clusters number and to propose a new function to assign the object at the most appropriate cluster. To measure the quality of a clustering and achieve a good data partitioning, several functions have been suggested: the entropy, the intra-class inertia and the inter-class inertia **[12] [13]**.
A good data partitioning is obtained by minimizing the entropy function, and by maximizing inter inertia (minimizing intra inertia). The proposed clustering model is based on the extraction of sequential rules from the sequential patterns technique.
In order to understand better the proposed approach, we present its principle as follows:



A. The Cleaning of the log file

The format of the web log files as presented in the form of a brute text file is unsuitable for a direct analysis by the various data mining techniques. The Cleaning and the structuring steps are necessary before any desired analysis. The preprocessing of the data consists to:

- **Pre-processing step**

The first step is to clean the log file by deleting:

- Unnecessary requests, namely Invalid queries whose status is less than 200 or greater than 399. This status indicates an informal message, a success, an error on the client's side or server's.
- Requests from web robots with unknown IP addresses, image requests and Scripts
- Requests whose method is different from <GET>. The GET method is a query in which the server processes the request and returns the contents of the object.

- **Data structuring step**

The second step consists to structure the log file to identify users and sessions:

- *Identification of sessions:* A session is composed of a sequence of requests that are sequentially ordered in time performed by the same user during a period of time.
  During this processing, the sequences composed of a single request are not considered and are eliminated in the pre-processing phase because they don't represent any interest for this analysis.
- *Identification of users:* The identification of the user can be done by his login and password, cookie, or by the IP address. We consider that two requests from two different IP addresses belong to two different sessions so they are performed by two different users. However, the inherent limitation to this method is the confusion produced between two different users using the same IP address can produce.

B. The generation of sequential patterns

To extract these patterns, several algorithms have been proposed which are based on time parameters such as the time window, the minimum and maximum duration that offers greater flexibility and significantly reduce the space of transactions. **[14] [15] [16]**.

Among these algorithms, there is: the Prefix tree algorithm **[4]**, the Generalized Sequential Patterns algorithm **[17]**, the Prefix Span algorithm **[18]**, the Sequential Pattern Mining algorithm, and Spade algorithm **[19]**.

In fact, we are interested in the Spade algorithm, for the following reasons:

- The Spade algorithm requires only a single reading of the database in order to represent the sequences as the lists of occurrences in the central memory.
- The Spade algorithm is based on the common prefixes of the sequences, so it groups together the sequential patterns by equivalence classes and thus decomposes the problem into sub-problems, where each equivalence class can be treated separately.
- Unlike the PSP and GSP algorithms which are level search algorithms, SPADE does not depend on I / O operations in the support count phase that triggers a read of the entire database.

These characteristics therefore reduce the memory space and the response time of the SPADE algorithm.

C. The generation of sequential rules

In the extraction of sequential rules from these patterns, ordering of events must be taken into account to reflect the order of the events that compose the rule. Based on this set of rules, the study of quality measures is essential to evaluate the quality of the generated rules and assess the association rules having the same interest. An extracted rule must satisfy the support and confidence measures which are mostly used as basic measures. The support measure remains a heavily used measure for algorithmic reasons; it's really difficult to do without the support measure during the exploration phase of the itemsets trellis whose role is to restrict the search space in the trellis of itemsets.

However, the confidence measure makes sense intuitively, which is a major advantage when presenting a set of rules to expert because the conditional probability of the conclusion knowing the premise makes reading more directly. In addition to the great interest in the support and confidence measures as extraction criteria, these functions have a concrete sense of comprehension and are fully absorbed by the non-specialist user.

However, both measures have some limitations since the algorithms that use them to generate a very large number of rules that are difficult to manage and many have less interest. The condition of the support removes the rules having a small support while some may have very strong confidence and can present a real interest. To sum up, the support and confidence measures appear insufficient to measure and to evaluate the quality of the generated rules. To overcome the weaknesses of these measures, it must consider other measures to evaluate the quality of rules. For this reason, several criteria to define a good quality measures have been proposed and subsequently different measures are available in the literature. Among these measures, there are: **[20];** *Lift; Recall*; *Conviction; Pearl*; *Φ coefficient*; *Pietetsky-Shapiro; Novelty*; *Centered confidence*; *Loevinger; Sebag-Schoenauer; Index of involvement.*

Among the criteria that a good measure must satisfy in a rule, we cite: a measure must have a concrete sense and must be easy to interpret. It must be sensitive to the appearance of counter-example and favours the appearance of examples of the rule and allow the user to tolerate some counter-example while maintaining the interest of a rule. It must be sensitive to the size of the data, as it should not lose its discriminatory power when the data size becomes large enough. It must also be used with a pruning threshold to remove all rules that do not interest the user and should be used with a pruning threshold.

None of these measures simultaneously satisfy all these criteria. This is why; the problem of research of quality measures to discover the most relevant rules remains largely open. The choice of measures specifically depends on the application domain and on processed data type.

D. The determination of clusters number

The partitioning algorithms select randomly a number of clusters and seek iteratively the optimal partition. These algorithms converge in a significant number of iterations. They start repeatedly to choose the distribution that provides a better partitioning of data where a considerable loss of time and of a large number of calculations. To overcome this limit, this approach is based on the extraction of sequential patterns to fix the number of clusters. The technique of sequential patterns can extract a set of item sequences, commonly associated over a period of time well specified and highlights the inter-transaction associations to generating a set of sequential rules.

As mentioned earlier several quality measures are mentioned which are used to identify identical rules in the same interest. Among all the properties (criteria) that a measure must satisfy, authors have judged according to web application domain the importance of some properties than others. This choice allows us to remove some measures as: implication and index Pietetsky-Shapiro measures because these measures depend on the data size. Both measures vary from increasing way with the size of the data and risk losing their discriminatory power when the size becomes large enough. In this case, the size of the treated web data should not intervene in the evolution of the function.

After this, each chosen measure allows to calculate interest rules. The approach proposes to regroup the rules which have the same interest (quality) represented by a measure value which characterizes the generated group. Each cluster will be characterized by a single value obtained by this measure as along with a number of rules to verify having the same interest. To refine and to reduce the number of rules, it is desirable to keep from this set just the non-redundant rules. The minimum non-redundant rules are the most useful and relevant rules, which must be presented to the user. A rule is redundant if it conveys the same information or a less general information that the information conveyed by another rule of the same utility and the same pertinence. An association rule **r** isn't a minimal redundant if there isn't another association rule **r'** having the same support and the same confidence, whose antecedent is a subset of the antecedent of **r** and the consequence is a superset of the result of **r**.

E. The proposition of a new similarity measure to assign an object

In the partitioning algorithms, the cluster is represented by its centroid. However the centroid of cluster in any way reflects the behaviour browsers from the point of view pages component the cluster which is an aligned sequence. For this reason, the new representative center is calculated by summing the occurrences of user's items belonging to the same cluster which indicates if the page is visited or not by the user. These sequences constitute the new representatives of groups.

The partitioning algorithms use the distance function (for example Euclidean, Jaccard, Manathan) to assign an object to the nearest center that proves unsuited to web data. For this reason, the proposed approach uses the similarity measure that is dedicated to sequential patterns which are:

- Ordered sequences of items and not of items.
- The positions of the itemsets in temporal order when calculating the similarity should be taken into account
- The number of common items in the sequence should be taken into account

By taking into consideration these criteria, the proposed similarity measure follows these steps:

- The order of itemsets when calculating the similarity is obtained through the verification of sequential rules that represent the itemsets sequences and therefore the apparition order of the items compared to the temporal identifier.
- The similarity measure reflects the relationship between the items constituting the web object.
- The number of common items obtained by the intersection between the sequences representing the object to classify and the sequence representing the cluster center.

Concerning the affectation of items to the most appropriate cluster, the first object is assigned to each cluster by verifying the most of its non-redundant rules to avoid getting the empty clusters at the end of the partitioning. The assignment of the remaining objects is achieved when respecting these criteria:

- The object must verify the maximum non-redundant rules of this cluster.
- The similarity measure between the object and the cluster center must be the greatest.

This algorithm stops when all objects are affected. At the end of its execution, each generated cluster is characterized by a set of sequential rules defined by the importance degree measuring their interest via the items set of this cluster.

Once the model is built, it will be necessary to evaluate the clustering result via the evaluation measures of generated clusters. The best retained partition will be the one that minimizes the intra inertia and the entropy and that maximizes the inertia inter. In the case where these criteria are not simultaneously checked in the same partition, we propose to calculate the difference between two partitions and choose the partition guaranteeing the maximum deviation based on the following heuristic: formula

V. Implementation

The k-medoids algorithm and the proposed algorithm are implemented in Java by using a log file which is exported from the university server of science and technology of Oran (Algeria) in 2013.
The initial log file contains 14500 navigations. The cleaning and pre-processing phase obtained a cleaned and processed log file of 9630 navigations. However, after the data structuring phase, we have identified 700 sessions, each of them represents a set of requests ordered sequentially in time.

F. Implementation Step

Once the log file is processed and ready to be used, the implementation process is based on three major elements

Mea-sure	Threshold	Number of clusters	Total number of generated rules	non-redundant number of generated rules	Number of iterations	Intra Inertia	Inter Inertia	Entropy	Execution time (Second)
Recall	Sup=0.2 Conf=0.5	9	219	204	2	2093,47	21982004,73	0,10	37,23
	Sup=0.3 Conf=0.5	8	122	118	2	2101,59	19543804,65	0,06	32,42
Lift	Sup=0.2 Conf=0.5	7	219	204	2	2003,28	34225073,22	0,04	35,60
	Sup=0.3 Conf=0.5	7	122	118	2	2103,27	17102877,22	0,04	31,87
Centred confidence	Sup=0.2 Con=0.5	8	219	204	2	2103,27	19548905,22	0,04	36,59
	Sup=0.3 Conf=0.5	6	122	118	2	2103,27	14656849,22	0,03	31,70
Pearl	Sup=0.2 Conf=0.5	5	219	198	2	2106,40	4877113,24	0,13	36,34
	Sup=0.3 Conf=0.5	5	122	117	2	2106,40	4877113,24	0,13	31,47
Conviction	Sup=0.2 Conf=0.5	9	219	183	2	2092,44	21976580,81	0,08	269,95
	Sup=0.3 Conf=0.5	8	122	112	2	2181,96	19531518,08	0,05	31,86
Φ coefficient	Sup=0.2 Conf=0.5	5	219	187	2	2106,40	4877113,24	0,13	37,67
	Sup=0.3 Conf=0.5	4	122	112	2	1053,97	1218126,83	0.15	30,89
Sebag	Sup=0.2 Conf=0.5	5	219	183	2	2106,40	7323141,24	0,1	36,56
	Sup=0.3 Conf=0.5	4	122	112	2	1053,97	2441140,83	0.11	31,50
Loevinger	Sup=0.2 Conf=0.5	6	219	192	2	2106,40	4877113,24	0,13	36,16
	Sup=0.3 Conf=0.5	6	122	115	2	2106,40	4877113,24	0,13	32,01
Novelty	Sup=0.2 Conf=0.5	18	219	204	2	2464.15	989471.26	0.35	36.391
	Sup=0.3 Conf=0.5	17	122	118	2	2545.24	77852.06	0.41	32812

TABLE I: Evaluation of the clusters quality measures by the proposed algorithm

generate the smallest inertia and entropy and the largest inter in a minimal execution time compared to other measures.

Comparing the generated partitions obtained by the Lift measure, Recall measure and Conviction measure, the best measure which guarantees the good data classification is the Lift measure. This measure verifies the quality measures of the clusters in a minimal execution time

H. Execution of K-medoids algorithm

For a more detailed experiment, the number of clusters is varied for each launching. This algorithm is executed several times to the same number of clusters by taking the average of the results represented in the **TABLE II**

According to the TABLE II, the best partition generated by the k-medoids algorithm is obtained by the number of clusters= 18 because it generates the smallest inertia and entropy and the largest inter comparing to other partitions.
Beyond the number of clusters 18, it can be seen a stability in the generated partitions for the evaluations measures. The k-medoids algorithm requires 7 iterations to segment the data.

Number of clusters	**Number of iterations**	**Intra Inertia**	**Inter Inertia**	**Entropy**	**Execution time (Second)**
2	7	10491.44	343558.85	1.36	654.66
3	7	11238.02	247067.78	2.58	668.45
4	7	9468.09	215383.68	1.92	689.37
5	7	642.75	207673.29	1.69	674.23
6	7	7258.16	246107.17	2.98	693.45
7	7	8416.24	260536.28	1.61	623.64
8	7	6204.27	231782.04	2.94	608.75
9	7	7698.37	289089.48	3.57	662.96
10	7	7150.82	98684.50	2,39	665.32
11	7	8801.87	317734.06	1.92	667.56
12	7	9301.80	332137.05	2.83	609.75
13	7	8866.82	929553.95	1.89	634.64
14	7	9197.96	1100838.55	1.82	634.45
15	7	7848.52	914427.19	1.98	607.99
16	7	10628.75	110007.38	2.70	623.33
17	7	8125.53	1086297.00	2.87	689.75
18	**7**	**7156.45**	**1198139.00**	**2.53**	**634.95**
19	7	7176.67	1172506.51	2.58	608.73
20	7	7158.06	1042908.67	2.59	691.94
21	7	7209.89	1069828.89	2.64	645.34
22	7	7243.17	1060007.42	2.68	644.65
23	7	7267.67	1058351.08	2.72	608.37
24	7	7212.55	1056506.52	2.78	612.37
25	7	[illegible]	[illegible]	[illegible]	[illegible]

TABLE II: Evaluation of the clusters quality measures by the k-medoids Algorithm

VI. Discussion

Comparing the best partition of the k-medoids algorithm and the proposed algorithm, the results show that the proposed approach verifies theses criteria:

- The inertia intra and entropy calculated by proposed approach are smaller than these calculated by k-medoids algorithm
- The inertia inter calculated by proposed approach is higher than that calculated by k-medoids algorithm
- The time classification calculated by proposed approach is less than that calculated by k-medoids algorithm

VII. Conclusion

This article proposes a new clustering model based on sequential rules in the Web Usage Mining. On the web data partitioning, the performance of any clustering algorithm is based on the similarity measure and on the number of determinate clusters that are the key to success of any clustering algorithm.

The proposed approach develops a new clustering algorithm which is based on the partitioning method of the clustering technique and also based on the Spade algorithm for the generation of sequential patterns. These patterns allow generating a set of sequential rules to be subsequently grouped according to their interest and evaluated through the quality evaluation measures of association rules.

To validate this approach, k-medoids algorithm and the proposed approach have been implemented for a comparative study of the performance according to the quality evaluation measures of obtaining clusters for each partition. This article compares the quality of obtained clusters of these algorithms and concluded that the proposed approach offers a better partition of data in a shorter running time.

References

[1] Lasek, P. "Efficient Density-Based Clustering", Unublished doctoral dissertation. University of technology WARSAW, Pologne, 2011.

[2] Eduardo Ares Brea, M. "Constrained Clustering Algorithms: Practical Issues and Applications". Unublished doctoral dissertation. University of Coruna, Spain, 2013.

[3] Liu, B. Book of Web Data Mining, Exploring Hyperlinks, Contents and Usage Data. Springer, Berlin, 2011.

[4] Masseglia, F. "An efficient algorithm for Web usage mining". Journal of Networking and Information Systems, 10, pp. 1-33, 2000.

[5] Norwati, M., Jalali, M. & Jalali, M. "Expectation maximization clustering algorithm for user modeling in web usage mining systems". European Journal of Scientic Research, 32(4), pp. 467-476, 2009.

[6] Jaskowiak, P. "On the Evaluation of Clustering Results: Measures, Ensembles and Gene Expression Data Analysis". Unublished doctoral dissertation. University of San Carlos, Cebu, Philippine, 2016.

[7] Malinen, M. "New Alternatives for K-Means Clustering". Unublished doctoral dissertation. University of Finland est, Finland, 2015.

[8] Anderson, C.R., Domingos. P, Weld.D.S, "*Adaptive Web Navigation for Wireless Devices*", pp. 879-884, 2001.

[9] M.S. Hacid, "Model-Based Cluster Analysis forWeb Users Sessions", ISMIS 2005, LNAI 3488, Springer-Verlag Berlin Heidelberg, pp. 219–227, 2005.

[10] Hay, B., Wets, G., Vanhoof, K"., *Mining navigation patterns using a sequence alignment method"*, Journal Knowl Information System, vol.6, pp.150–163. 2004.

[11] Poornalatha,G., Raghavendra Prakash,S., "*Web sessions clustering using hybrid sequence alignment measure (HSAM)"*, Springer-Verlag, 2012.

[12]Rosenberg, A., & Hirschberg, J. "V-Measure: A conditional entropy-based external cluster evaluation measure". Proceeding of the international conference in Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 410–420, 2007.

[13] Koskela, M., Laaksonen, J., & Oja, E. "Entropy-based measures for clustering and SOM topology preservation applied to content-based image indexing and retrieval". Proceeding of the 17th International Conference on Pattern Recognition, 2004.

[14] Kam, P. & Wai-chee Fu, A. "Discovering temporal patterns for interval based events". Journal of Data Warehousing and Knowledge Discovery, pp.317-326, 2001.

[15] Méger, N. "Recherche automatique des fenêtres temporelles optimales des motifs séquentiels ». Unublished doctoral dissertation. University of Lyon, France, 2004.

[16] Saneifar, H., Bringay, S., & Maguelonne Teisseire, A. « S2MP: une mesure de similarité pour les motifs séquentiels ». Proceeding of the international conference in Evaluation des méthodes d'Extraction de Connaissances dans les Données, pp. 35-46, 2009.

[17] Ayres, J. "Sequential patterns mining using a bitmap representation". Proceeding of the 8th international conference in knowledge Discovering and Data mining, 2002, pp.429-435.

[18] Pei, J., Han, J., Pinto, H., & Mortazavi, B. "Prefix-Span: Mining sequential patterns efficiently by prefix projected pattern growth". Proceeding of the 17th international conference in Data engineering, pp. 215-224, 2001.

[19] Zaki, M.J. "SPADE: An Efficient Algorithm for Mining Frequent Sequence". Journal of Machine Learning, 42, pp.31–60, 2001.

[20]Feno-Rajaonasy, D. « Mesures de qualité des règles d'association normalisation et caractérisation des bases ». Unublished doctoral dissertation. University of Reunion, France, 2007.