Comparative Study of Quality Measures of Sequential Rules for the Clustering of Web Data

Hadj-Tayeb Karima[1] , Belbachir Hafida[2]
[1,2]Computer science Department, Sciences and Technology University-USTO-Oran-Algeria,
[1]k.hadjtayeb@Gmail.com
[2] h-belbach@yahoo.fr

***Abstract*—To exploit large databases in the Web, data mining techniques have been applied. Among these techniques, the cluster analysis and the extraction of sequential patterns are considered to be the most important aspects in the process of exploring the web to find large groups.**
Web data that we handle are streams of sequential data where time plays a vital role in sequential patterns found to extract sequential rules. In this case, the ordering of events must be taken into account in the measure of calculation in order to measure the quality and interest of a rule.
The purpose of this study is to construct a model of clustering based on the grouping of sequential rules by quality measures. We aim at the end of our study to detect a good measure of applicable data quality and provide a good partitioning through the measures evaluation of the clustering quality.

Keywords—clustering, sequential patterns, sequential rules, quality measures, Web data, measurement devaluations clusters

VI. INTRODUCTION

The important growth of information available on the internet requires tools to search for more efficient and effective strategies to discern relevant information from hundreds or thousands of page views that can be structured by an analysis of web users.
To understand better the behavior of browsers and satisfy their needs, it is imperative to process and analyze these data by applying data mining techniques such as: association rules, classification, clustering and sequential patterns. The goal is to discover hidden relationships between users and useful as well as between users and Web objects and consequently improve the performance of web services.

Among these techniques, cluster analysis and sequential patterns can be considered as the most important aspects in the process of Web Mining which is a stream of sequential data when time is in the essence of the sequential rules extracted.
In the case of extracted rules from sequences, the scheduling events in the calculation of the measurement should be taken into account. The measures derived from conventional measures, support and confidence are most used to characterize the sequential rules. However, the excessive use of these two measures is not sufficient to ensure the quality of detected rules. As part of this study, there is a large number of measurements to characterize the association rules alongside the choice of a measure depend largely on the scope and criteria that the measure must satisfy. In the context of sequential rules, measures require a non-trivial adaptation to reflect the order of the events that make up the rule.

Our aim behind this study is to create a model for clustering web users based on sequential patterns and clustering rules through quality measures. At the end of our study, we expect to detect a good measure of quality that guarantees a good partitioning of our data in terms of assessing the quality of clustering and computation time measurements

VII. CLUSTERING TECHNIQUE [1] [2]

Clustering may be defined as a set of methods used for cutting a set of objects into groups (clusters) based on the attributes that describe. The goal of clustering is to understand how to group objects in the same cluster observations to be similar according to some metric (homogeneity intra-class) and place comments deemed dissimilar in separate clusters (inter-class heterogeneity). Good data classification must optimize a criterion based on the inertia in the goal to minimize the intra-class inertia or maximize inter-class inertia.
A good clustering method ensures high similarity intra-group and low similarity inter-group dependent grouping criteria used by the method. In the literature, the clustering is based on two main approaches: the hierarchical approach and the approach of partitioning

A. Hierarchical Method

These methods are gradually classed hierarchically, ie, a tree which is called a dendrogram. Algorithms based on this method are trying to create a hierarchy of clusters; the most similar objects are grouped in clusters at lower levels, while the least similar objects are grouped

in clusters at the highest levels. In fact, there are two subtypes: agglomeration and division

B. Partitioning Method

The partitioning data is used to divide a series of data in different homogeneous clusters. Its principle is to divide the set of individuals in a number of classes by using an iterative optimization strategy. Subsequently, the general principle is to generate an initial partition, and then try to improve it by reallocating data from one class to another. Unlike hierarchical algorithms that produce a class structure, the partitioning algorithms produce one partition which leads to seeking local maxima in optimizing an objective function which reflects the fact that individuals should be similar within the same class and dissimilar from one class to another.

VIII. SEQUENTIAL PATTERNS IN THE WEB [3]

- A transaction is for a user C, a set of items representing all visited page views by C on the same date. In a database, a transaction is written as a triplet: <id customer-id-date, itemset>. An itemset is a non-empty set of items noted (i1i2 ... ik), where ij is an item.
- A sequence is a non-empty ordered list of itemsets denoted <s1s2 ... sn> where sj is an itemset (a sequence is a series of transactions with an order relation between transactions). A data sequence is a sequence representative visits a browser.

  We consider streams of sequential data in the web, or T1; T2 ... Tn transactions ordered by date and is growing itemset (Ti) all items corresponding to Ti, then the data sequence is <itemset (T1) itemset (T2) ... itemset (Tn)>.

IX. STUDY OF ALGORITHMS SEQUENTIAL EXTRACTION: REASONS FOR THE GENERATION OF RULES SEQUENTIAL [4] [5]

In the literature, several algorithms have been proposed; we briefly introduce the pioneer GSP algorithm, the PSP algorithm, the algorithm Spade and PREFIX-SPAN algorithm

A. GSP Algorithm (Generalized Sequential Patterns)

This algorithm starts by sorting the initial database based on the unique identifier as the primary key and CID as a secondary key. The use of the identifier of this time base is to transform it into a sequence data base, and it is the latter which is analyzed by the algorithm. After making the first point of the sequence-based data to determine the set of frequent sequences, the GSP generates all k-candidate sequences of step k from the (k-1) frequent sequences step (k-1) by performing the join of F (k-1) with itself, called self-join of F (k-1).

The algorithm alters the phase between generation of the candidate sequences and calculation phase carriers generated sequences to determine the common points among them. It is based on the hash tree to represent the candidate sequences that will be stored in the leaves.

B. PSP Algorithm (Prefix tree for Sequential Patterns)

The PSP algorithm provides a data structure prefix tree to represent the candidate sequences or any path from the root to a node of the tree represents a single candidate. Moreover, any candidate sequence is represented by one and only one path of the root to a node. The candidates' generation of length 2 is similar to GSP, by cons for the higher levels, PSP pulls profile the structure of the prefix tree as follows: for each leaves of the tree, PSP Research the root item represented by x. Then, it stretches the sheet for building these copies of son of x. In this step, the algorithm applies a filter for only generating sequences that it knows in advance that they cannot be frequent.

C. SPADE algorithm

This algorithm performs a single reading of the sequence database to represent it in the main memory in a form of sequence of occurrences, for all subsequent treatments will be made on these lists. To generate the candidate sequences, SPADE offers to subdivide the space research equivalence class. The candidates generation of length (k + 1) is performed by temporal joins between two; all the lists are frequent occurrences k-sequences belonging to the same equivalence class k, ie those sharing the same prefix length of (k-1). The calculation of the support of candidates is to verify the cardinality of occurrences obtained lists and keep only the frequent sequences

D. PrefixSpan Algorithm (Prefix Projected Sequential Pattern mining)

This algorithm is proposed to reduce the number of generated sequences. By exploiting like previous algorithms common prefixes that often present data sequences. However, its strategy is much different to the extent that it does not generate any candidate sequence during different phases of the research. The algorithm performs successive projections of the base sequence data for the partition based on common prefixes. In its first phase, it identifies all frequent items (1-prefixes), and it builds intermediate bases which are projections of the latter on each frequent 1-préfix, which built the second and final round of base sequences. The algorithm seeks to grow the length of sequential patterns using this method recursively.

X. GENERATION OF SEQUENTIAL RULES AND MEASURES OF QUALITY RULES [6]

The generation of rules is much less expensive than the generation of frequent patterns since it is no longer necessary to the expensive route of the database. To generate the rules, we consider the set F of frequent patterns found in the previous phase. From these frequent subsets, we can generate all the valid rules in the context of data mining their respective Trusts exceed the minimum threshold of minimal support and confidence.

Be the rule in the form: x ⟶ y

A. Support: It is defined by

$$\text{Supp}(X \rightarrow Y) = p(X' \cap Y'). \quad (1)$$

It indicates the proportion of entities verifying both the premise and the conclusion of the rule. It is a symmetric measure and takes values between [0, 1]

B. Trust: It is defined by

$$\text{Conf}(X \rightarrow Y) = p(Y'|X') = \frac{p(X' \cap Y')}{p(X')}. \quad (2)$$

It says that the proposed entities give satisfying conclusion among those checking the premise of the rule. It is not sensitive to the size of data. Accordingly, It is a non-symmetric measure and takes values between [0, 1].

The number of valid association rules in the sense of a measure of quality is often very high which creates a new problem for the user to know the difficulty of assessing the value of extracted rules. It is in this context that the quality measures have been proposed in order to quantify and rank the association rules.

There is several quality measures proposed in the literature, the most used are probably the support and trust. However, these measures can generate a very large number of rules that are very difficult to manage and many of which have little interest what makes these two inadequate steps ensure the quality of the rules. To overcome these weaknesses, several measures have been proposed to check several criteria, namely:

C. Recall: It is defined by

$$Rappel(X \rightarrow Y) = p(X'|Y') = \frac{p(X' \cap Y')}{p(Y')}. \quad (3)$$

This measure assesses the proportion of entities satisfying the premise among those which satisfy the conclusion of the rule. It is insensitive to the size of the data. It is a non-symmetric measure in which case measure takes values between [0, 1]

D. Lift: It is defined by

$$Lift(X \rightarrow Y) = \frac{p(X' \cap Y')}{p(X')p(Y')}. \quad (4)$$

This represents the ratio of independence between the premise and the conclusion of the rule. It is a symmetric and sensitive measurement data size. It takes values between

[0, + ∞ [

E. Conviction: It is defined by

$$Conviction(X \rightarrow Y) = \frac{p(X')p(\overline{Y'})}{p(X' \cap \overline{Y'})}. \quad (5)$$

It indicates that the number of examples against the rule is less than that expected by the assumption of independence between the premise and conclusion. It is a non-symmetric measure and takes values between [0, + ∞ [

F. Pearl: It is defined by

$$Pearl(X \rightarrow Y) = p(X')|p(Y'|X') - p(Y')|. \quad (6)$$

This measure is used to evaluate the interest of a rule with respect to assumption of independence between the premise and conclusion. It is a symmetric measure and takes values between [0, 1]

G. Pietetsky-Shapiro: It is defined by

$$Piatetsky(X \rightarrow Y) = np(X')(p(Y'|X') - p(Y')). \quad (7)$$

It assesses the interest of a rule from its deviation from independence. It is symmetrical, and sensitive to the size of data. It takes values between [-n, n]

H. trust-centered: It is defined by

$$\text{Conf}_{\text{centrée}} = p(Y'|X') - p(Y'). \quad (8)$$

It allows taking into consideration the size of the conclusion and measures the influence of achieving the conclusion by contribution more than the premise. It is sensitive to non-symmetrical and the size of data. It takes values between [-1, 1]

I. Loevinger: It is defined by

$$Loevinger(X \rightarrow Y) = \frac{p(Y'|X') - p(Y')}{p(\overline{Y'})}. \quad (9)$$

It standardizes the measurement confidence centered by the number of entities that do not verify the conclusion. It is sensitive to non-symmetrical and the size of data. It takes values between] -∞, 0 [

J. Reduced contraction: It is defined by

$$Contramin(X \rightarrow Y) = \frac{p(X' \cap Y') - p(X' \cap \overline{Y'})}{p(Y')}. \quad (10)$$

It evaluates the difference between the numbers of examples against a ruler. It selects the rules with more examples than against examples. It takes values between

] -∞, + ∞ [

K. New: It is defined by

$$\text{Nouveauté}(X \rightarrow Y) = p(X' \cap Y') - p(X')p(Y'). \quad (11)$$

It measures the deviation from independence between the premise and the conclusion of the rule. It is symmetrically dependent on the size and data. It takes values between [-1, 1]

L. Sebag: It is defined by

$$Sebag(X \rightarrow Y) = \frac{p(X' \cap Y')}{p(X' \cap \overline{Y'})}. \quad (12)$$

It evaluates the ratio between the number of examples and examples against the rule. If the value is greater than 1, the rule has more than an example against such. It is not symmetrical and takes values between [0, + ∞ [

M. Degree of involvement: It is defined by

$$Ind\text{-}Implication(X \rightarrow Y) = \sqrt{n}\frac{p(X' \cap Y') - p(X')p(Y')}{\sqrt{p(X')p(Y')}}$$

(13)

It estimates the number of cons example relative to the expected under the assumption of independence quantity. It is not symmetrical and varies depending on the data size. It takes values between [-√n, + ∞ [

XI. CLUSTERING WEB USERS [7] [8]

Clustering of users in the field of web browsing sessions grouping, the web developer can help to better understand the browsing behavior of users to provide their personalized services most suited to their needs as quickly as possible. Therefore, understanding how visitors use the Web site is one of the essential steps of website developers that will implement intelligent Web servers in real time to be able to dynamically adapt their designs to meet the needs of future users

This work explores the concept of Web Usage Mining from web session is represented as a sequence characterized by the IP address of the browser, the visited pages and the date of each page. We propose in this work a new algorithm for clustering data represented web users based on frequent sequential patterns

A. *The log file for web data*

With the popularity of the WWW, very large amounts of data such as address or user requested URLs are automatically collected by Web servers and stored in files access log. A log file is used to collect data by servers that represent the database web sequential. Each entry in the log file represents a request made by a client machine to the server.

A log is a set of entries in the Access log file. An entry G belonging to Log, is a tuple:

$$g = < ip_g, \{(l_1^g.URL, l_1^g.time), ..., (l_m^g.URL, l_m^g.time)\} > \quad (15)$$

Such that for 1 <= k <= m, lgk.URL represents object requested by the browser g to date lgk.time and for all

1 <= j <k, lgk.time> lgj.time

B. *Disadvantages of clustering approaches*

Under the data clustering, the methods mentioned above were the main limitation of being dependent on baseline (initial centers) representing clusters defined previously. They build partition k clusters of base D of n objects and gradually permit more refined classes and therefore can give the better classes. In fact, the algorithms need to run multiple times with different initial states to obtain a better outcome by following each iteration the reallocation mechanism that reallocate points between classes. Each initialization (set number of clusters) corresponds to a different solution (local optimum), which can in some cases be far from optimal. A naive solution to this problem is to run these algorithms multiple times with different initialization and retain the best combination found. The use of this solution is limited due to its high cost in terms of computation time and the number of steps for the best score can be obtained after repeated execution of the algorithm.

XII. CLUSTERING APPROACH PROPOSED WEB BASED DATA MINING SEQUENTIAL PATTERNS

To overcome the limitations of clustering methods, we rely on sequential patterns to establish our classification model uses data from the web. Among the extraction of sequential patterns algorithms presented above, we looked at Spade algorithm for the following reasons:

- It requires only one reading the database to represent the sequences as lists of occurrences in the main memory
- It is based on common prefixes of sequences, so the group sequential patterns by equivalence classes and thus breaks down the problem into sub problems to be addressed in memory which reduces the memory space
- Unlike the PSP and GSP are search algorithms by level algorithm, SPADE does not depend on I / O operations in the phase count of the support which triggers a reading of the entire database

Consequently, these features reduce the response time of the algorithm SPADE.

As part of the proposed quality measures, there is a large number of measurements to characterize the association rules and the choice of a measure depends largely on the scope and criteria that the measure must satisfy. In the case of association rules derived from sequences, the scheduling

events should be considered in the calculation of the measurement. Both measures, derived from traditional measures of association rules used to characterize the rules sequential measures are support and confidence. The algorithms using these measures generate a large number of rules that are very difficult to manage and many of which have little interest. Then, the condition of support that drives the extraction process removes the rules with little support while some may have a very high confidence and can have a real interest. Finally, the exclusive use of quality measures and Trust Support not enough to guarantee the quality of the rules detected.

To overcome these problems, the measures described above have been proposed. As part of our study based on web sequential data, we exclude the measure: novelty, Degree of involvement, Pietetsky-Shapiro as its measures depend on the size of the data while we process large web data where the size should not be intervened in the evolution of the function. In addition to this reason, the new measure does not satisfy the condition that the measure must tolerate little against examples to keep the interest of the rule. Once patterns are extracted, the set of rules is generated and will be evaluated to understand better the value of every extracted rule. Based on all of the quality measures we propose to consolidate the rules of associations with the same interest (quality) represented by a measured value.

The goal is to build a model of optimal classification adaptable to our database. For this purpose, we propose to make a comparative study of quality measures and generate only offer a further better quality of classification

After getting our clusters, data classification is based on the verification of the entire rule, knowing that any association rule is as**: if premise then conclusion.**

XIII. MEASURES OF QUALITY ASSESSMENT OF A CLUSTERING [9]

To assess the quality of clusters obtained after partitioning of the data, three steps are calculated

A. Entropy:

Is a measure of quality to measure how different classes of objects are divided into a cluster, such that:

- The entropy of a cluster C nr size is calculated using the following formula:

$$E(C) = -\frac{1}{logq}\sum_{i=1}^{q}\frac{n_r^i}{n_r}log\frac{n_r^i}{n_r} \quad (16)$$

Where q is the total number of clusters and n_r^i is the number of sequences of the ith cluster that are part of the cluster C.

- Entropy Clustering is then given by the formula:

$$Entropie = \sum_{r=1}^{k}\frac{n_r}{n}E(C_r) \quad (17)$$

Where n is the total number of sequences.

We consider a small entropy value which indicates a good clustering with respect to the reference clustering.

There are other steps to get a good score which is to minimize intra inertia and maximize inter inertia:

B. Intra Inertia

- The inertia of a intra-cluster measures the concentration of points cluster around the center of gravity is calculated by:

$$jk = \sum_{i\in ck} d^2(xi, uk) \quad (18)$$

u: is the center of gravity

$$uk = \frac{1}{Nk}\sum_{i\in ck} xi \quad (19)$$

Total Inertia intra partitioning is the summation of inertia within clusters. More inertia is low, the smaller the dispersion of the points around the center of gravity

C. Inter Inertia

- Inertia -inter a cluster measures the distance from the centers of the clusters together. It is calculated by:

$$jb = \sum_{k} N_k d^2(uk, u) \quad (20)$$

u is the center of gravity:

$$uk = \frac{1}{Nk}\sum_{i\in ck} xi \quad (21)$$

The total inertia inter partitioning is the summation of inertia inter clusters. More inertia, the larger clusters are well separated in order to get a good score

XIV. IMPLEMENTATION

Our goal is to propose an approach based on frequent patterns approach to exploit various information relatives on the use of a website. The goal is to classify browsers of this site recorded in the log file in minimal time.

We implemented this algorithm on a platform java log file test 1000 records over a period of 4 days. We first perform preprocessing of the log file that was done in two phases:

A. Phase 1

- Removal of unnecessary queries since their appearance does not reflect any behavior relating to the Internet that is invalid queries, requests for images and multimedia files type the extension (jpg,. wma ...), the scripts usually downloading and requested by a user and leave the final urls that reflect the Web

Mea-sures	Nb cl	Nb reg	M_V	intra	Ent-ropy	Tp
trust centered	11	6	-26, -23 -22, -20 -17, -16	448.55	9.34	5266
Pearl	8	6	320,67, 106,94 100,52, 181,124	418	10.4	4875
Convictio n	6	6	0,1,2, 3,4,6	435.61	9.73	5110
Reduced contractio n	**2**	**6**	**-1, -2**	**287.76**	**2.88**	**4600**
Loevinger	**2**	**6**	**-1, 0**	**287.76**	**2.88**	**4562**

pages assigns the extensions page: html, htm, php, ...

- Delete records with a "post" method

B. Phase 2

We applied the Spade algorithm in order to extract frequent patterns then extract sequential rules. We subsequently consolidated these rules and classified our browsers in clusters so defined.

For a comparative study of quality measures on the same data set, we combined our rules according to several quality measures for each measure and describe all clusters found subsequently to classify all our browsers.
The algorithm stops when the number of iterations is reached or achieves stability.

At the end of our study, we must maximized the inertia inter or minimized the inertia intra. We calculated each defined measure:

- Nb_cl: represents the number of clusters generated for each calculated measure
- Nb_reg: represents the number of sequential rules for all generated cluster
- M_V: represents the measurement value obtained by combining rules
- Intra: represents the total inertia of all partitions
- Entropy: represents the total entropy of all partitions.
- Tp: Represents the calculated execution time Mili-Second
- nb_iter: represents the number of reached iterations.

- **To min support = 2**

After several values of iterations, we set, the maximum number of iterations to 5 and we fix the minimum support = 2 and minimal confidence 0.33
We launched the algorithm on measures and Recall Lift and Sebag. We found that it generates only a single metric value and therefore only one cluster in which case they were excluded for this data.

Mea-sures	Nb cl	Nb reg	M_V	Nb iter	intra	Ent-ropy	Tp
Pearl	4	4	320, 117, 181, 104	6	565.62	11.3	5234
trust-centered	3	4	-26, -23, -22	2	285.72	4.34	4984
Convicti on	3	4	2, 4, 6	2	284.22	4.34	5219
Loeving er	**2**	**4**	**-1, 0**	**2**	**280.48**	**2.78**	**4800**

TABLE I. Results of this study for supp=2

- **To min support = 3**

We launched the algorithm on measures Reduced contraction
Lift, Sebag, Recall . We found that it generates only a
single metric value and therefore only one cluster in
which case they were excluded for this data.

TABLE II. Results of this study for supp=3

XV. DISCUSSION

The best classification for a sup = 2, is the one that minimizes the intra and provides a small entropy value, for these reasons, we choose for our sample data, Reduced contraction and Loevinger measures because they ensure the best clustering in minimum execution time
The best classification for a sup = 3, we choose the measure Loevinger as it ensures the best clustering in minimum execution time
After comparing the results of the two tables, Loeviner measure is selected for support = 2 and confidence = 0.33 because:
• It generates for the same number of cluster = 2, a smaller

number of rules than that obtained by support= 3

• It executes in a smaller number of iterations than that obtained by support = 3

• It provides a value of entropy and intra better than that obtained by support = 3

• It runs in a smaller execution time

XVI. CONCLUSION

We presented in this paper a comparative study of quality measures for grouping of associations rules.

Our job is to build a model of clustering based on sequential patterns and clustering of sequential extracted rules in order to categorize Web data.

To this end, we conducted a comparative study of measures of quality of association rules to detect good quality measure applicable to our data and provide a good partitioning through the evaluation measures of the quality clustering in a minimum execution time.

We found at the end of our study for the same parameters input and the same sample data, the measure Loevinger meets the criteria initially namely the evaluations measures of the quality of obtained clusters

REFERENCES

[36] A. Belhabib, O.Lagha, « Développement d'une application à base de l'algorithme de classification k-means ». Telemcen, Algérie, 2012.

[37] S.Guha, R.Rastogi, and k. Sim, « Cure : an efficient clustering algorithm for large databases », SIGMODD, 1999.

[38] A.Ben Zakour, « Extraction des utilisations typiques à partir de données étérogènes historisées en vue d'optimiser la maintenance d'une flotte de véhicules », Thèse Doctorat, université BORDEAUX I, 2012

[39] J.pei, J.Han, H.Pinto and B.Mortazavi, "Prefix-Span: Mining sequential patterns efficiently by prefix projected pattern growth. 17th international conference on data engineering (ICDE'01), Heidelberg, April 2001

[40] M.Zaki, "Spade: An efficient algorithm for mining frequent sequences", Machine Learning Journal, Vol 42 (1-2), pp 31-60, January 2001

[41] D.Rajaonasy FENO, « Mesures de qualité des règles d'association :normalisation et caractérisation des bases »,Thèse Doctorat en Informatique, Université de La Réunion, 2010

[7] S.Guha, R.Rastogi, and k. Sim, « Cure : an efficient clustering algorithm for large databases », SIGMODD, 1999.

[8] A.GOMES DA SILVA, « Analyse des données évolutives : application aux données d'usage du Web », Thèse Doctorat Université Paris IX 2010

[9] B.Liu, « Web Data Mining, exploring hyperlinks, contents and usage data», Springer verlag berlin. 2011.