3rd International Conference on Automation, Control, Engineering and Computer Science (ACECS'16)
Proceedings of Engineering & Technology (PET)
pp. 796-801
Copyright IPCO-2016

Extraction of Grammatical Rules Based on Sequential Mining

Asma Ksiksi[1], Chekib Gmati[2], Hamid Amiri[3]

Signal, Images and Technologies of Information laboratory (LR-SITI)

National Engineering School of Tunis (ENIT)

Tunis, Tunisia

asma.ksiksi@gmail.com, chekibgmt2007@yahoo.fr, hamidlamiri@gmail.com

Abstract—**the sequential association rule mining is an important technique to extract an interesting knowledge from database. In this article we will use this technique to extract Arabic grammatical rules and to provide the relation between the unites of sentence. We will exploit the area of a sequential rule to extract a meaningful one that give a significant information about the relation between the different items.**

***Keywords*— Frequent pattern mining, Sequential association rules mining, Arabic grammar, Arabic text.**

I. INTRODUCTION

With the massive rise in the information's volume these days, and the emergence requirements for a superior technique to access to this information, there has been a strong resurgence of interest in data mining research [1]. It consists in the extraction of implicit and hidden information from a huge amount of data.

One of the most important and well researched techniques of data mining for the extraction of knowledge in large database, is the association rules. It intendes to extract frequent patterns, interesting correlations, associations or casual structures between the sets of items in the transaction databases [3]. If we introduce the notion of order between items, the extraction should be based on this notion, in order to extract a sequential association rules.

In this paper we propose to use sequential association rules to extract grammatical rules. Firstly, we will present the complexity of arabic language and the grammatical properties. Then we will introduce the notion of association rules and the sequential association rules. Finally we will explain our approach and presented the obtained results.

II. ARABIC LANGUAGE

Arabic language is used by millions of people in north Africa and in twenty Middle East countries. It is the fifth spoken language in the world [8] [9]. It represents the most contemporary spoken Semitic language today with more than 300 million speakers [10].

This language is distinguished from the other one by a wealth morphological and syntactic system. It is flexional, excessively derivational, and agglutinative language; in fact Arabic morphological system can generate many words originating from the same radical (root) but they're not necessarily semantically convergent.

The construction of lexical units in the Arabic language and their transformation in accordance with the desired meaning is so varied. However, the grammatical coherence is ensured by analyzing the units' position and case marking ;for example the adjective «النَّعْت» must agree with the nouns in terms of number, gender, state of definiteness and grammatical case. Also for the definite noun «إسم معرّف », the name must be preceded by the definite article ال.

Several references exist on the Arabic morphology. For example, [11] and [12] realized a detailed description of the grammar and morphology of Arabic language. More recently, a description of the principal syntactic structures of the Arabic language was developed by [13]. He involves an in-depth study on the dependences between words, the order, the agreement, and the syntactic disambiguation problem. [14]

Due to the increasing number of Arabic content on the Web, the need for specialized tools to analyze and understand Arabic text has emerged. [15] suggested an automatic approach of construction of ontology by using the method of "repeated segment" to identify the relevant terms that indicate the concepts associated with the domain then they used the "co-occurrence" to link these new concepts to the ontology by hierarchical or nonhierarchical relations.

Also the authors of [16] proposed a model for representing Arabic knowledge in the Computer Technology domain using Ontologies. They combined the traditional and the modern Arabic words to serve semantic based search and retrieval of Arabic blogs on the Web.

Among the most known resources, we cite the ARABIC WORDNET [17]. In the article [18] authors proposed to improve the performance of this resource by using a morpho-lexical patterns to add semantic relations between synsets. These patterns are extracted from Arabic wikipedia corpus, composed of 2050 articles. They obtained a set of 135 morpho-lexical patterns.

MADAMIRA [26]., is a system for morphological analysis and disambiguation of Arabic language that combines some of the best aspects of two previously commonly used systems for Arabic processing, 2007) MAda+ Token [23] and AMIRA [24].

ISSN: 2356-5608

ASMA [27], is a system for automatic segmentation and morphosyntactic disambiguation of Modern Standard Arabic. The system ASMA performs both inflectional morpheme segmentation and agglutinative clitic segmentation.

Article [19] involves the problem of detecting and correcting hidden semantic spelling errors in Arabic texts. This approach is based on the multi-agent architecture: a group of syntactic agents to treat syntactical anomalies and a group of semantic agents to detect semantic inconsistencies in the sentence.

Moreover, there are a variety of problemes of automatic extraction of relationships from Arabic text treated by ontology. In [20], authors proposed to implement an enhanced version of Hearst's algorithm and to integrate it into a framework. This evaluation takes three directions : first of all, the use of a recall metric for measuring the correctness of extracted patterns with respect to existing correct ones' then the use of a precision metric for measuring the ability of their proposed methodology to detect patterns with respect to all retrieved information. Finally applying the F-measure that denotes the overall accuracy.

Among the techniques of extraction rules, we find the association rules mining that provide the association between items in database.

III. Association Rules Mining

An Association rule mining is a famous knowledge discovery technique for finding association between the items from a transaction database.It represent an implication of the form A→B where A and B are an itemsets. The sets of items A is called antecedent and B consequent of the rule. It provides information about the existing relations between the items A and consequent B. It expresses how objects or items are related to each other, and how they can be grouped together.

The first step to extract the association rules mining is to find out the frequent itemset which is called candida=t_e items [2].

This transaction can be measured in terms of its support and confidence.

The support sup $(A \rightarrow B)$ is defined as the relative frequency of transactions in the data set D that contains the itemset A and B.

$$\mathrm{Sup}(A \rightarrow B) = \mathrm{Sup}(A \cup B) = \frac{|\{t \in D : A \subseteq t \text{ et } B \subseteq t\}|}{|D|} \qquad (1)$$

The confidence conf(A -> B) of a rule measure the reliability of the inference given by rules.

$$\mathrm{Conf}(A \rightarrow B) = \frac{\mathrm{Sup}(A \cup B)}{\mathrm{Sup}(A)} \qquad (2)$$

Then, the important association rules are filtered from the candidate itemsets.

A rules r is available only if sup (A -> B) > minsup and conf(A -> B) > minconf where minsup represents the threshold of support and minconf represents the threshold of confidence. These two values are specified by user.

The strength of the approach mining association rules is the exponentially number of rules that can be extracted from a variety of data set such as texts, images, sounds …

For textual data, the Association rules mining are used to solve many difficult problems in natural language processing.

Article [2] has proposed a method based on the Association rules in order to remove any ambiguity on the word sense. They have used Apriori algorithm to extract the association rules between the sense of the ambiguous words and contexts. But this method doesn't consider the temporal dimensions that can give more useful descriptions to assign the appropriate mining.

The authors of [4] make a survey about the problem of Word sense disambiguation based on association rules to determinate the sense which the most association rules deduced.

In [5], Chao Tang has presented a new model for grammatical rules mining from Chinese text. This model proceeds in three steps; pre-processing, building association rules mining and verification of her availability. The experimental result shows that the algorithm works better on small sentences.

IV. Sequential Patterns Mining

A sequential pattern is represented as a sequence of itemsets that occurs sequentially with a specific order. This ordered elements or events are stored with or without a concrete notion of time.

The problem of sequential pattern mining was initially introduced by Agrawal and Srikant in [6]: given a set of sequences, where each sequence composed of a list of elements organized by transaction time and each element is a set of items, the objective of the sequential pattern mining is to detect all of the frequent subsequences in order to predict a plausible occurrence frequency in the set of sequences where the support of patterns is higher than or equal to the min support threshold specified by user. [7]

As we consider the hierarchies of concepts and the hierarchies of relationships, the number of test patterns can be many times higher than the number of patterns to test, if we consider only the objects [21].

Sequential association rules can be more informative and beneficial than frequent association rules especially to detect grammatical rules in the Arabic sentence.

V. Motivation

The extraction of grammatical rules in Arabic language is a complex task because this language is characterized by a complex morphological and syntactic system. In our work we will propose a method to extract interesting rules by using sequential association mining.

The association rules ensure the extraction of the most frequent patterns and the relations between terms of sentence to obtain grammatical rules. But to know that a sequence appear frequently is not sufficient to predict a meaningful grammatical rules.

In Arabic language to extract the grammatical coherence between unites of Arabic sentences, it's important to consider the notion of order. Since classical association rules do not consider the chronological order of the different objects, another category of approaches is used: it consists in sequential patterns which take into account the time aspect.

The use of sequential rules provides the extraction of relations between the items in the order given in the sentence. It ensures the extraction of a meaningful grammatical rule.

Depending on the choice of the thresholds minconf and minsup, current algorithms can provide an extremely large amount of results, omitting valuable information. In our work, the dimension of antecedent and consequent affects the meaning of grammatical rules.

VI. Proposed Approach

Our work consists on the extraction of association rules in order to find relations between the properties whose characterize grammatical rules while considering the order of items.

To explore the search space of frequent sequential rules, we will vary the minsup threshold with the aim to extract the most meaningful rules.

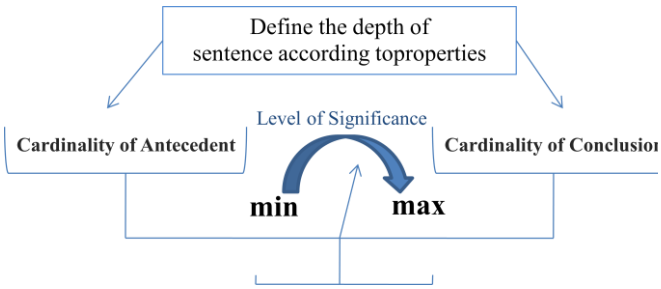

Fig. 1 Variation of the cardinlity of antecedent/ conclusion

The Fig 1. shows that it is necessary to find some balance between antecedent and consequent to extract the most significant rules. The cardinality of antecedent and conclusion has an impact on the enrichment of grammatical rules consequently on the validity of rules.

Also it is important to mention that this equilibrium is related to the quality of association rules that depend on values of minsup. In our worsk we will highlight the relation between the threshold minsup and the cardinality of antecedent and consequent.

The analysis of this equilibrium, taking into account the minimum threshold minsup, will lead us to recognize the association rules valid in our case. In fact, it will serve to highlight one of the hidden characteristic of Arabic grammatical rules. This characteristic consists in the appropriate structure of the rule that makes it understandable and valid. Through the properties of association rules we can give prominence to this feature. In other words we will highlight the relations between the proprieties or between items, frequency and cardinality between the antecedent and conclusion. In fact, this property (cardinality of the antecedent and conclusion) reflects the depth of the grammatical rule in the sentence (number of words from right to left). This is necessary because the complexity of the Arabic language is very high, in point of view the sentence structure. If we analyze the Arabic language, then it is essential to take into account the nature of language and its structural proprieties .

VII. Experimental Evaluation

During the first phase, we do a morphological analysis using the open source analyzer Al Khalil Morpho Sys. Our choice was made on Al Khalil MorphoSys because it can treat non diacritic texts and the texts partially or totally diacritized. Alkalil is based on the modeling of a wide set of Arabic morphological rules and the integration of language resources that are useful to the analysis.

Alkhalil Morpho Sys linguistic resources are composed of over than 250 million words from eight Arab corpus available online including contemporary and ancient books. The morphological Analyzer Alkhalil Morpho Sys is considered one of the most important open source Analyzers. [25]

In 2010, it has won the first position among the 13 Arab morphological systems around the world in a competition organized by the Arab League Educational, Cultural and Scientific Organization (ALECSO).

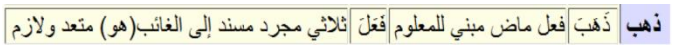

Fig. 2 Result after morphological analysis by the analyser Alkhalil

We used a corpus of 500 sentences in the literary field. The first step is to extract the proprieties of the terms of each sentence by applying Khalil, considering the order of words and meaning (from right to left).

Every sentence is a sample that has several properties which we will call items. Then we assign for each properties an integer (i.e 1.2...) in the order of the sentence structure: Prefix → Word type→Part of Speech (POS)→Suffix; and respecting the order of words in the sentence. So, we have for each sample a set of items that respects the morphology of the sentence. Thus, a database is built, on which we will apply the algorithms for discovering sequential rules mining.

In our analysis, we manipulated the cardinality values of antecedents and conclusions by performing cross combinations, taking into account the variations in the value of minsup.

There are many algorithms related to sequential rules mining, in our work we will use the ERMiner (Equivalence class based sequential Rule Miner). It represents a novel algorithm developed on 2014 based on searching equivalence classes of rules having the same antecedent or consequent. ERMiner is characterized by a fast execution time even with a huge database [22]. For this reason we choose this algorithm.

We performed experiments to ensure the extraction of meaningful rule using the ERMiner algorithm with the evaluation of a linguistic expert. According to an expert, a grammatical rule is meaningful if it has some number of properties. Example of meaningful rule:

فعل ماضي مسند إلى الغائبة (هي) ت تاء التأنيث الساكنة --> إسم جامد مفرد مؤنث تاء التأنيث

Example of meaningful rule:

إسم جامد مفرد في حالة التعريف ---><مصدر أصلي منصوب

To discover the influence of the cardinality of antecedent and consequent, we will select the thresholds minsup and minconf, and vary the number of antecedent and consequent.

However, we note that the results are similar if we change the value of the minconf parameter so we choose to vary just the minsup parameter. The minconf is fixed to 0,5.

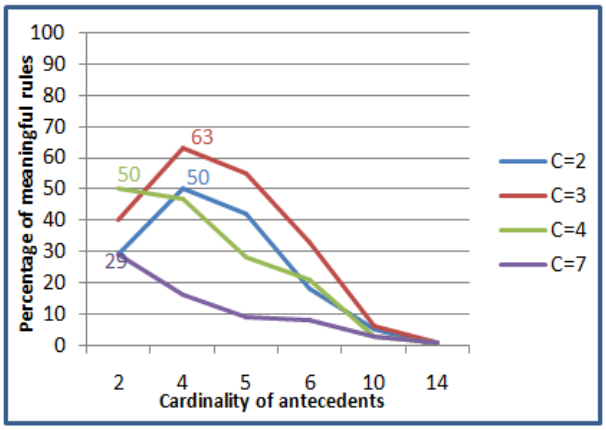

Fig. 2 The variation of meaningful rules

We ran the ERMiner algorithm with minsup=0,3 and minconf=0,5. The observation shows that to extract highest number of meaningful rules, the number of antecedent and conclusion must be considerable and reflects some depth at the sentence. According to Fig.2 we note that the cardinality values of the conclusion and cardinality values of antecedent: {4→(2,3,4)},{2→4}{5→(2,3)}, give to us the best results. for four antecedent and three consequents we obtain the best values of meaningful rules which is 63%. For four antecedent and two consequents we obtain the best values of meaningful rules 50%. For four antecedent and two consequents we obtain the best values of meaningful rules 50%. For two antecedent and four consequents we obtain the best values of meaningful rules 50%.

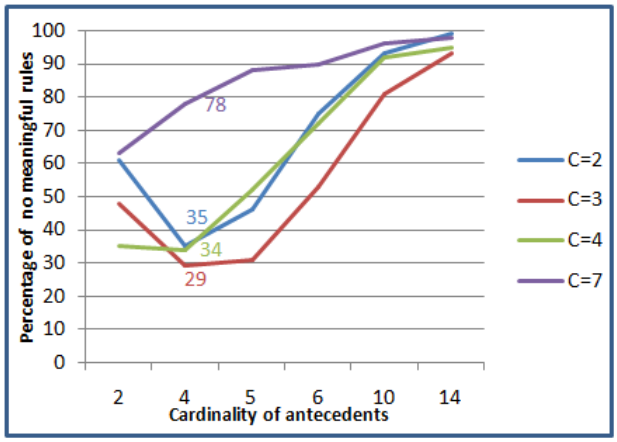

Fig. 3 Variation of No Meaningful Rules

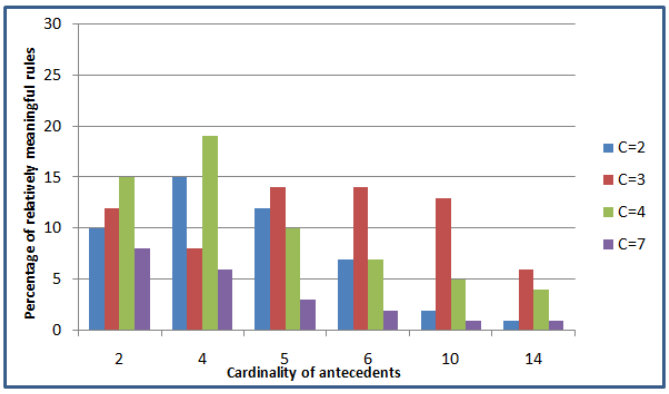

Fig . 4 Variation of Relatively Meaningful Rules

In Fig.3, the recognition rates of invalid rules, confirms the results of Fig. 1 regarding both of the cardinalities {4→(2,3,4)}, which have respectively the rates 35%, 29% and 34% and the cardinalities {5→(2,3)}, which have respectively the rates 35%, 31% et 46%.

The recognition rates of the relatively meaningful rules are also proportional to the results of Fig.1.

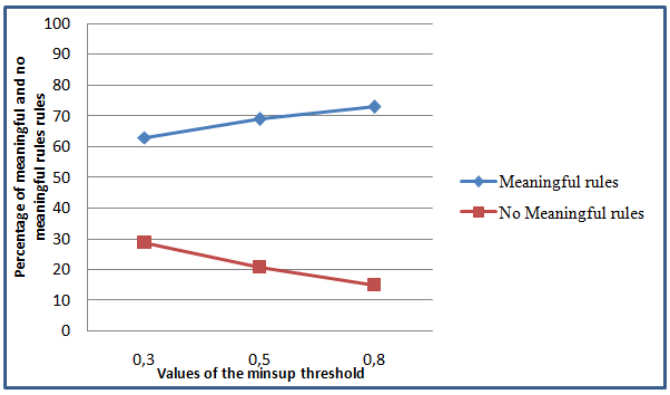

Fig 5: Variation of the minsup threshold

The percentage of extracting meaningful rules changes by varying the values of minsup. While the minsup threshold decreases until the number of significant rules decreases consequently the number of insignificant rules becomes higher. In Fig 5, we notice that the rate of valid rules depends on the increase in values of minsup threshold. This indicates conclusively that the extraction of association rules represents adequately the Arabic grammatical rules.

VIII. CONCLUSION

In this paper, we proposed to use the sequential association rules to extract Arabic grammatical rules. This technique allowed finding the relations between the items of sentence. So, it ensured the extraction of a meaningful grammatical rule.

To provide most significant rules, we explored the search space of frequent sequential rule. . The analyses have leaded us to extract the assembly of items into blocks and to take into account the order between these sets of items. Therefore, we will direct to the sequential extraction of association rules through researching sets of frequent items.

REFERENCES

[1] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, 3nd ed, Morgan Kaufmann publishers: Elsevier, 2011, p.14, 15

[2] Y. Sun and K. Jia, *Research of Word Sense Disambiguation Based on Mining Association Rules*, Intelligent Information Technology Application Workshops, (2009), pp. 86-88

[3] Agrawal, R., Imielinski, T., and Swami, A. N. 1993. *Mining association rules between sets of items in large databases.* In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, P. Buneman and S. Jajodia, Eds. Washington, D.C., 207{216.

[4] Samit Kumar, Neetu Sharma, S. Niranjan ,*Word Sense Disambiguation Using Association Rules: A Survey*, International Journal of Computer Technology and Electronics Engineering. 2012;2(2)92-98

[5] C. Tang and C. Liu, Method of Chinese Grammar Rules Automatically Access Based on Association Rules, in Proc. Computer Science and Computational Technology (ISCSCT 2008) vol.1, 2008, pp. 265 – 268.

[6] AGRAWAL R. & SRIKANT, R*Mining sequential patterns.* In Int. Conf. on Data Engineering : IEEE ,1995.

[7] CELLIER P., CHARNOIS T. & PLANTEVIT M. *Sequential patterns to discover and characterise biological relations*. In Computational Linguistics and Intelligent Text Processing, LNCS, p. 537–548 : Springer, 2010.

[8] M. Paul Lewis, *Ethnologue: languages of theworld*. Dallas, Texas: SIL International, pp. 5, 2013.

[9] W. Chung, *Web searching in amultilingual worl*, Magazine:Commun. ACM- Web searching in a multilingual, vol. 51 Is 5,pp. 32–40, 2008.

[10] N. Habash, *Introduction to Arabic natural language Processing. Synthesis Lectures on Human Language Technologies*, Morgan &Claypool Publishers , pp. (15–17, 21, 27,29, 56), 2010.

[11] Kouloughli, DJ. E. *Grammaire de l'arabe d'aujourd'hui*, Paris : press pocket, (1994).

[12] Wtson, *The phonolgy and morphology of arbic* , The phnology of the word's languages, Oxford linguistics, 2008

[13] M. Attia et al, "Improved spelling error detection and correction for arabic",pp. 103-112, 2012.

[14] Souhir Gahbiche-Braham, *Amélioration des systèmes de traduction par analyse linguistique et th´ematique : application `a la traduction depuis l'arabe theses,*2013

[15] Boudabous, M.M.; Kammoun, N.C.; Khedher, N.; Belguith, L.H.; Sadat, F., *Arabic WordNet semantic relations enrichment through morpho-lexical patterns*, Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on, 1(6), pp 12-14, 2013

[16] Lilac Al-Safadi, Mai Al-Badrani and Meshael Al-Junidey. Article: Developing Ontology for Arabic Blogs Retrieval. International Journal of Computer Applications,19(4), pp 40-45, April 2011.

[17] W. Black, S. Elkateb, P. Vossen, "*Introducing the arabicwordNet project*", In Proceedings of the third International WordNet Conference (GWC-06), 2006.

[18] S. Carpendale et al, "Ontologies in biological data visualization". Computer Graphics and Applications, IEEE , vol. 34, no.2, pp. 8,15, 2014.

[19] C. Zribi, H. Mejri and B. Mohamed, "Un analyseur hybride pour la détection et la correction des erreurs cachées sémantiques en langue arabe". Traitement Automatique des Langues Naturelles : à Toulouse. 2007.

[20] Mohammed G.H. Al Zamil, Qasem Al-Radaideh, Automatic extraction of ontological relations from Arabic text, Journal of King Saud University - Computer and Information Sciences, 26(4), pp 462-472, 2014.

[21] Lisa Di Jorio, Lylia Abrouk, C´eline Fiot, Maguelonne Teisseire, Daniele herin, Enrichissement d'ontologie basé sur les motifs s´equentiels, theses, 2007.

[22] Philippe Fournier-Viger, Ted Gueniche1, Souleymane Zida, Vincent S. Tseng , *ERMiner: Sequential Rule Mining using Equivalence Classes*, Advances in Intelligent Data Analysis XIII: 13th International Symposium, 2014

[23] HABASH Nizar, RAMBOW Owen, and ROTH Ryan, "Mada+ tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization," in Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, 2009, pp. 102-109

[24] DIAB Mona, "Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking," in 2nd International Conference on Arabic Language Resources and Tools, 2009.

[25] Bebah, M.O.Abdallahi, A.Meziane, M.Azzeddine, and A.Lakhouaja. Alkhalil morpho sys. In 7th International Computing Conference in Arabic, May 31- June 2, 2011, Riyadh, Saudi Arabia.

[26] A.Pasha, M.Al-Badrashiny, M.Diab, A.Kholy, R.Eskander, N.Habash, M.Pooleery, O.Rambow, and R.M.Roth, "*MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic*", LREC, 2014

[27] M. Mageed, M.Diab, S.Kubler, "ASMA: A System for Automatic Segmentation and Morpho-Syntactic Disambiguation of Modern Standard Arabic", RANLP, 2013.