Computer-Assisted Reading: Getting Help from Text Classification and Maximal Association Rules

Ismaïl Biskri
Université du Québec à Trois-Rivières / Laboratoire de Mathématiques et Informatique Appliquées, Canada
Email : Ismail.Biskri@uqtr.ca

Abdelghani Achouri, Louis Rompré, Steve Descoteaux and Boucif Amar Bensaber
Université du Québec à Trois-Rivières / Laboratoire de Mathématiques et Informatique Appliquées, Canada
Email : {Achouri, Rompre, Descoteaux, Bensaber}@uqtr.ca

***Abstract*—The combination of text classification and maximal association rules will allow the extraction of hidden knowledge, often relevant from the text and allow the detection of dependencies and correlations between the relevant units of information (words) of different classes. In fact, the results of text classification take the form of large and noisy classes of similarities.**

***Index Terms*—Classification, Maximal association rules, Computer assisted reading**

I. Introduction

In recent years, a phenomenal growth in the amount of text information is noticed. Because of the rise of the Web or because of electronic documents in various institutions, computer-assisted reading has become of major importance. We may need computer-assisted reading as a preliminary to various human tasks such as: text's summarizing, text's topic identification, information retrieval, or intratextual and intertextual navigation in large corpus.

Text classification has been the focus of interest of many experts of computer-assisted reading for a long time [1]. It seems that it brings out useful co-occurrence patterns for computer-assisted reading. The main goal of text classification is to group into "homogeneous" classes textual objects that share similar properties [2] [3]. The result is a set of classes of similarity usually displayed as lists of words (in fact called bag of words) that co-occur together. These classes can sometimes seem less significant or completely insignificant. They are often very large and despite many improvements, they are very noisy. The process of the maximal association rules extraction downstream of a classification operation is an interesting avenue to enable the discovery of relevant lexical associations for an informed decision. The M-support and the M-confidence are two discriminating measures that expert user may consider to « clear » classes of similarity and accelerate the interpretation by assisting the user in computer-assisted reading.

Among text classification algorithms, the best known include: Knn, Kmeans, ART, SOM, SVM. Classifiers are based on the common principle of a vector representation of documents with using a matrix of frequencies (possibly the presence / absence) of units of information in each document. The main goal of text classification is to group into "homogeneous" classes textual objects that share similar properties. The result is a set of classes of similarity usually displayed as lists of words (in fact called bag of words) that co-occur together. These classes can sometimes seem less significant or completely insignificant. They are often very large and despite many improvements, they are very noisy. Some classes share, also a part of their lexicon. This constitutes a major obstacle to an objective interpretation of the extracted knowledge made by a human. We believe that it is necessary to develop tools to facilitate the interpretation of classes, and thus, to enhance the interest of the classification.

We believe, moreover, that the identification of maximal associations can play a major role in computer-assisted reading and so in practical applications as information retrieval, construction and maintenance of ontologies, etc.

II. Maximal Association Rules

A brief survey of the literature on data mining [4] teaches us that association rules allow for a representation of regularities in the co-occurrence of data (in the general sense of the term) in transactions, regardless of their nature. Thus, data that regularly appear together are structured in so-called association rules. An association rule is expressed as $X \Rightarrow Y$. This is read as follows: each time that X is encountered in a transaction, so is Y. There are also ways to measure the quality of these association rules: the measure of Support and the measure of Confidence. Other measures of quality of associations rules are proposed in the literature and many studies are

Manuscript received April 24, 2013; Copyright credit, Project number, corresponding author : Ismaïl Biskri.

© 2013 ACADEMY PUBLISHER
doi:10.4304/jait.4.4.157-165

dedicated to their evaluation [5] [6]. Among existing measures support and confidence are the most common.

The concept of association rule emerges mainly from the late 60 [7] with the introduction of the concept of the support and the confidence. Interest in this concept was revived in the 90s through the work of Agrawal [8] [9] on the extraction of association rules in a database containing business transactions. Currently, work is being done on how best to judge the relevance of association rules, as well as the quality of their interpretation [6] [10] [11], and their integration into information retrieval systems [12] and into classification processes for text mining [11].

To illustrate association rules, consider the definition of the principal elements in the following example:

- Three transactions to regroup the data that co-occurs: T1:{A, 1, K}; T2:{M, L, 2}; T3:{A, 1, 2}
- Two sets to categorize the data: E1:{A, M, K, L}; E2:{1, 2}
- X and Y: two separate sets of information units: X:{A}; Y:{1}. X $\subseteq$ E1 and Y $\subseteq$ E2.

For a transaction Ti and a set of information units X, it is said that Ti supports X if X$\subseteq$ Ti. The Support of X, noted as S(X), represents the number of transactions Ti such that X $\subseteq$ Ti. In the case of transactions T1, T2 and T3, S(X) = S(A) = 2.

The Support of the association rule X $\Rightarrow$ Y is the number of transactions that contain X and Y. In the case of our example S(X$\Rightarrow$ Y) = S(A$\Rightarrow$ 1) = 2.

The Confidence of the association rule X $\Rightarrow$ Y, noted as C(X $\Rightarrow$ Y), corresponds to the support of this association rule divided by the Support of X otherwise stated as C(X$\Rightarrow$ Y) = S(X$\Rightarrow$ Y)/S(X). In the case of our example, C(X$\Rightarrow$ Y) = C(A$\Rightarrow$ 1) = 1.

Despite their potential, association rules cannot be established in the case of less frequent associations. Thus, certain associations are ignored since they are not frequent. For example, if the word *printer* often appears with the word *paper* and less frequently with the word *ink*, it is very probable that the association between *printer* and *paper* will be retained to the detriment of the association between *printer, paper* and *ink*. In fact, the confidence criterion associated to the relationship between *printer, paper* and *ink* would be too low.

The maximal association rules, noted as $X \xrightarrow{max} Y$, compensate for this limitation. They are dedicated to the following general principle: each time that X appears alone, Y also appears. Note that X is reputed to appear alone if and only if for a transaction Ti and a category set Ej (X $\subseteq$ Ej), Ti $\cap$ Ej = X. In this case, X is maximal in Ti with regards to Ej and Ti M-Supports X. Note the M-Support of X by Smax(X), which thus represents the number of transactions Ti that M-Support X.

In the transaction T1, X is not alone with regards to E1 since T1 $\cap$ E1 = {A, K}. On the other hand, in the transaction T3, X is alone since T3 $\cap$ E1 = {A}.

The M-support of the maximal association $X \xrightarrow{max} Y$ noted as Smax($X \xrightarrow{max} Y$) represents the number of transactions that M-support X and support Y.

In the case of our example, only the transaction T3 M-supports X while T1 and T3 support Y. Consequently Smax ($A \xrightarrow{max} 1$) = 1.

The M-confidence noted as Cmax($X \xrightarrow{max} Y$) represents the number of transactions that M-support $X \xrightarrow{max} Y$ relative to the set of transaction that M-support $X \xrightarrow{max} E2$. The M-confidence of the rule $X \xrightarrow{max} Y$ is thus calculated by the formula Cmax($X \xrightarrow{max} Y$) = Smax($X \xrightarrow{max} Y$)/ Smax($X \xrightarrow{max} E2$). In the association $A \xrightarrow{max} 1$, the M-Confidence is found to be equal to 0.5.

Finally, it should be noted that we must define the minimum thresholds for the M-support of a maximal association, as well as for its M-Confidence.

III. Identification of Maximal Association Rules in Similarity Classes

GRAMEXCO (n-GRAMs in the EXtraction of knowledge (COnnaissance)) is our prototype that has been developed for the numerical classification of multimedia documents [13], particularly text documents. The numerical classification takes place by way of a numerical classifier.

The unit of information considered in GRAMEXCO is the n-gram of characters, the value of *n* being configurable.

The main objective is to provide the same processing chain, regardless of the corpus language, but with easily legible layouts in the presentation of the results. Recall that the use of n-grams of characters is not recent. It was first used in work by Damashek [14] on text analysis and work by Greffenstette [15] on language identification. The interest in n-grams today has been extended to the domains of images, and musicology, particularly in locating refrains [16]. A character n-gram is defined here as a sequence of *n* characters: bigrams for n=2, trigrams for n=3, quadrigrams for n=4, etc. For example, in the word *computer* the trigrams are: *com*, *omp*, *mpu*, *put*, *ute*, *ter*. Even if there is no theory to guide the choice of the optimal unit of information [17], we justify our choice of n-gram of characters as the unit of information by: (i) The cutting into sequences of n consecutive characters is possible in most languages. It is necessary that any approach can be adapted to several languages because of the "multilingual" nature of the web; (ii) The necessary tolerance for a certain ratio of deformation or flexion of lexical units [18]. The functioning of GRAMEXCO is not entirely automatic. The choice of certain parameters is made by the user according to their own objectives. GRAMEXCO takes a raw (no indexed) text as input in UTF format. There are then three first main steps where the user can customize certain processes.

© 2013 ACADEMY PUBLISHER

The **first step** consists of building a list of information units and information domains (parts of texts to be compared for similarity). From the two operations carried out simultaneously, we retrieve an output matrix with a list of the frequency of appearance of each information unit in each information domain. The information units may be in the form of bigrams, trigrams, quadrigrams, etc. Obtaining information domains passes through the process of text segmentation which may be done in words, phrases, paragraphs, documents, web sites or simply in sections of text delimited by a character or a string of characters. The choice of the size of the n-gram and the type of textual segment is determined by the user according to the goals of their analysis.

The **second step** consists of reducing the size of the matrix. This operation is indispensable given the important cost in resources that an overly large matrix would represent. During this step, a list of n-grams undergoes some trimming that corresponds to:

(i) the elimination of n-grams whose frequency is lower than a certain threshold or above another threshold,
(ii) the elimination of specific n-grams selected from a list (for example, n-grams containing spaces or n-grams containing non-alphabetic characters),
(iii) the elimination of certain n-grams considered as functional, such as suffixes.

In the **third step**, the classification process takes place. The classifier used here is chosen between Fuzzy-ART [19], K-means [20], SOM [21]. The choice of classifier is not dictated by particular performance reasons since this is not our objective. We could have just as easily chosen another classifier that would have admittedly yielded different results. Such variations continue to be the focus of research such as was presented in Turenne [22].

At the end of this step, segments considered as similar by the classifier are regrouped into similarity classes. Furthermore, the lexicon of these segments forms the vocabulary of the classes to which they belong.

The classes obtained at the end of the classification operation will be the transactions of the process that will allow the extraction of maximal association rules [23]. Finally, in order for the process to be carried out, it must be supervised by the user who will have to first determine the word for which the most probable associations will be found.

To illustrate this step, let us posit the following scenario that will allow us to discover maximal association rules $X \xrightarrow{max} Y$ based on the results of a classification.

The input of the classification is a text in which the vocabulary represents a category set E1: {x, a, b, c, d, e, f}. The classification outputs classes with their respective lexicon: C1 : {x, a, b, c}, C2 : {a, c, d}, C3 : {x, e, f, d}.

If the classes represent the transactions, the vocabulary of the input text represents a set E1 for categorizing the textual data (the vocabulary) in which set X is chosen.

This being established, the extraction process of maximal association rules is carried out in three steps:

1st step: choice of set X: it is the user who chooses the lexicon from a list of elements of E1 that will represent X. In our case X and E1 coincide. Let us assume for explanatory purposes that X = {x}.

2nd step: identification of set Y and set E2: the identification of the category set E2 in which Y would be a subset largely depends on the set X selected and on the classes of which X is a subset.

In the case of our illustration, X is included in C1 and in C3. Y may therefore be a subset either of {a, b, c} or of {e, f, d}. In other words, Y may represent one of the following subsets: {a}, {b}, {c}, {a, b}, {a, c}, {b, c}, {a, b, c}, {e}, {f}, {d}, {e, f}, {e, d}. {f, d}, {e, f, d}.

The measures of M-Support and of M-Confidence will be calculated with regards to these different possible values of Y. An iterative process would allow for testing the set of these possibilities. We may, however, limit the number of iterations in order to avoid an overly prohibitive computational cost, for example, by fixing (via parameter) the cardinality of subset Y.

Let us suppose that Y = {a, c}; in order to construct E2, the respective categories of elements a and c must first be established. These are obtained by uniting classes that contain a (or c, respectively). Consequently, E2 = category(Y) = category{a, c} would be obtained by intersecting category(a) with category(c). Thus:

$$category(a) = \{a, b, c\} \cup \{a, c, d\} = \{a, b, c, d\}$$

and

$$category(c) = \{a, b, c\} \cup \{a, c, d\} = \{a, b, c, d\}$$

therefore :

$$E2 = category(Y) = category(a, c) = category(a) \cap category(c) = \{a, b, c, d\}$$

3rd step: once the sets E1, E2, X and Y as well as the transactions have been clearly identified, the calculation of the measures may be made.

Consider the association $x \xrightarrow{max} a, c$. Using the classes C1: {x, a, b, c}, C2: {a, c, d}, C3: {x, e, f, d} as transactions, and E2 = {a, b, c, d}, it follows that M-support equals 1, since only Class 1 contains X= {x} and Y= {a, c}, and an M-confidence of 0.5 since two classes contain X while only one contains X and Y.

© 2013 ACADEMY PUBLISHER

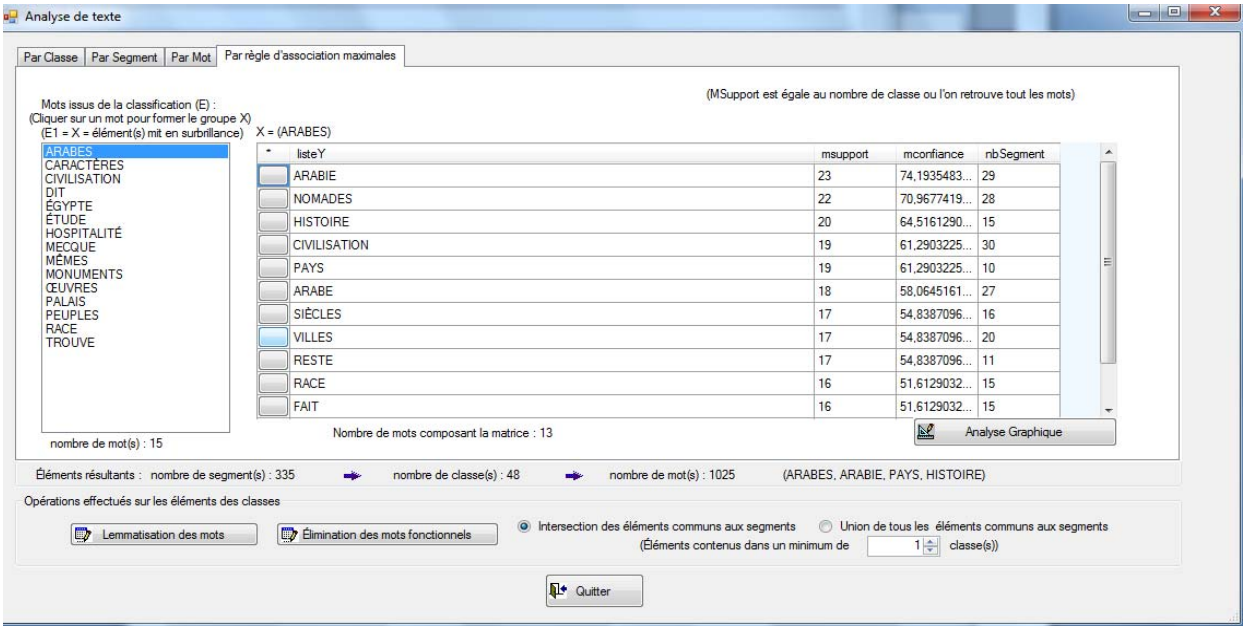

Figure 1: First interface for displaying results

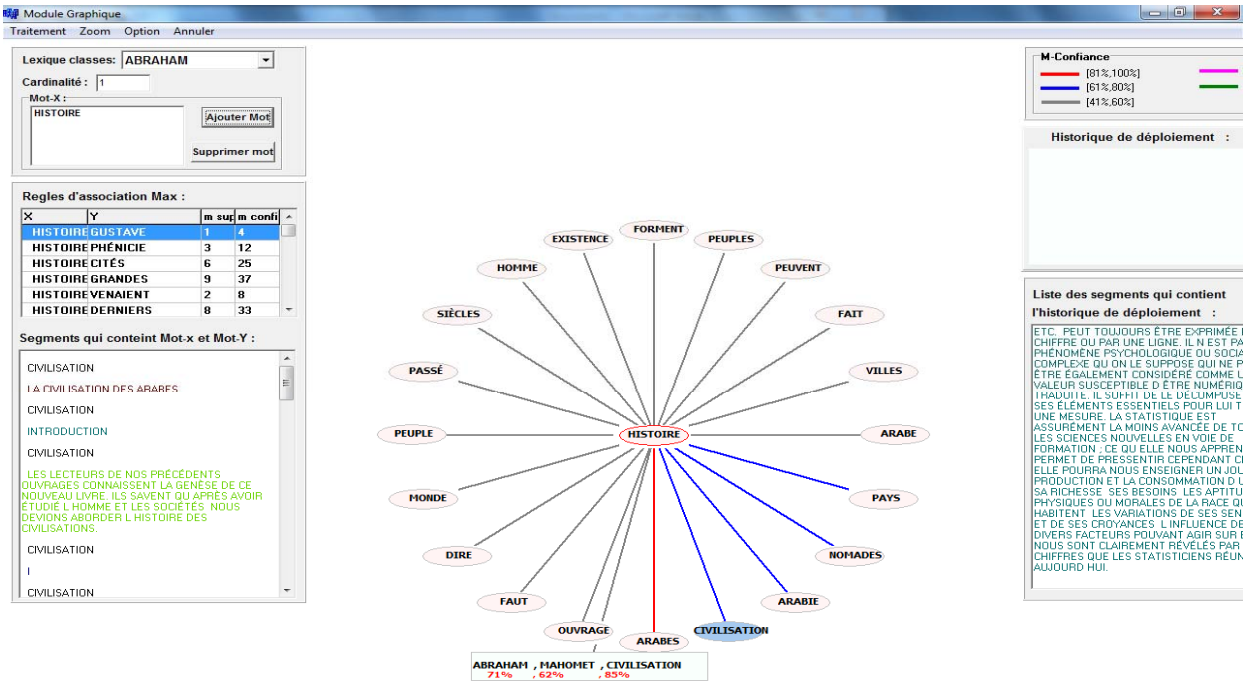

Figure 2: Second interface for displaying results

The whole of the theory presented here was implemented in C#. The results of the analyses are stored in XML databases. The software has several interfaces including, particularly two, which allows the display of results. In the first of these two interfaces (Figure 1), the results take the form of a list of words Y that co-occur with a selected word X in a lexicon representing the vocabulary of the text to analyze, according to their M-support measures and M-confidence measures. In the second interface (Figure 2), a graphical representation provides an overview of the extracted maximal association rules for one selected X.

IV. Experiments

We show the following first four experiments to illustrate the maximal rules extraction. The experiments were applied to four corpora (three of them are extracted

© 2013 ACADEMY PUBLISHER

from web sites). Two corpora are in French and two are in Arabic. The first corpus is a collection of interviews with directors of small and medium Quebecois businesses in order to learn about their perspectives on the notion of *risk*. The second corpus addresses the history of the reign of *King Hassan II* of Morocco. The third corpus (in Arabic) addresses the *Organisation of the Petroleum Exporting Countries* (OPEC). Finally, the fourth and final corpus (in Arabic) summarizes the biography of the American President, *Barack Obama*. The domains are sufficiently different to draw conclusions on the efficacy of the methodology. Note: we limit ourselves to show just maximal associations and scores of each association (M-support and M-Confidence). We assume that the reader is sufficiently familiar with the methods of classification and we do not need to show classes of similarities.

1st experiment: the corpus, as mentioned above, addresses the perspective of directors of small and medium Quebecois businesses with regards to the notion of *risk*. One of the constraints during the interviews was the obligation put on the directors to use the word *risk* when they deemed it necessary. In our experiments, this aspect is crucial since we need to know which words are associated to *risk* in the discourse of the directors.

Thus, despite the presence of noisy data such as, for example, *Pause* and *x*, which were intentionally inserted into the text for ethical reasons (*x* represents the name of people who were questioned) and to represent silences (*Pause*), interesting results were still obtained. For example:

- *Risk* $\xrightarrow{\max}$ *Project* is an association that is found in 10 classes (M-support = 10) with a confidence of 100%.
- *Risk* $\xrightarrow{\max}$ *Management, Project* is an association that we find in 7 classes (M-support = 7) with a confidence of 70%. In other words, 30% of the time, it is possible to find the word *Risk* in classes where *Management* and *Project* did not occur together.
- *Risk* $\xrightarrow{\max}$ *Management* is an association that we find in 7 classes (M-support = 7) with a confidence of 70%.
- *Risk* $\xrightarrow{\max}$ *Product* is an association that we find in 5 classes (M-support = 5) with a confidence of 50%.

The following table summarizes the results obtained:

TABLE I.
RESULTS OF THE 1ST EXPERIMENT

Y	M-Support	M-Confidence
Client	1	10%
Shareholders, Cost	1	10%
Client, Project	1	10%
Decision, Product	2	20%
Year	2	20%
Markets, Price	2	20%
Science	3	30%
Interview, Studies	3	30%
Function	4	40%
Manner, Level	5	50%
Product	5	50%
Question	6	60%
Interview, Risk	6	60%
Level, x	7	70%
Management	7	70%
Management, Project	7	70%
Project, Risks	8	80%
X	10	100%
Pause	10	100%
Project, x	10	100%
Pause, x	10	100%
Project	10	100%

2nd experiment: For the second experiment, we chose a short 4-page text about the reign of *King Hassan II*. For this experiment, we intentionally chose to consider the cardinality of set Y equal to 1. For X = {*Hassan*}, we obtained the results summarized in table 2.

Note that, for example, the association *Hassan* $\xrightarrow{\max}$ *II* is very strong. Its confidence is 100%. Likewise for the associations *Hassan* $\xrightarrow{\max}$ *Morocco* and *Hassan* $\xrightarrow{\max}$ *King*. Although their confidence is only 61.54%, this is sufficiently high to consider the two associations as maximal.

3rd experiment: For the third experiment, we chose an Arabic text regarding the *Organisation of the Petroleum Exporting Countries* (OPEC), the goal being to evaluate the validity of the method with regards to the Arabic language. For the purposes of the experiment, we chose X = {OPEC}. The table 3 provides a summary of the results (a translation of the Arabic words is provided):

© 2013 ACADEMY PUBLISHER

TABLE II.
RESULTS OF THE 2ND EXPERIMENT

Y	M-Support	M-Confidence
Doctor	1	7.69 %
Professor	1	7.69 %
Spain	1	7.69 %
Tunisia	1	7.69 %
Spanish	2	15.38 %
Journalist	3	23.08 %
History	3	23.08 %
Prepare	3	23,08 %
Title	4	30,77 %
France	5	38.46 %
Politics	6	46.15 %
Year	7	53,85 %
King	8	61.54 %
Morocco	8	61.54 %
II	13	100 %

TABLE III.
RESULTS OF THE 3RD EXPERIMENT

Y	M-Support	M-Confidence
Mechanisms	1	9,09 %
Paris, Countries	1	9,09 %
Creation, prices	2	18,18 %
Petroleum	3	27,27 %
Countries, members	3	27,27 %
Prices	3	27,27 %
Organisation, prices	3	27,27 %
Creation	3	27,27 %
Members	4	36,36 %
Summit	4	36,36 %
World	4	36,36 %
Organisation, country	4	36,36 %
Organisation	6	54,55 %
Countries	7	63,64 %
In	9	81,82 %

The results obtained indeed show the tight relationship between the acronym *OPEC* and the two words *Organisation* and *Countries*. However, there is an association with a relatively high M-support and M-confidence that relates *OPEC* to the function word *in*. We consider this association as being noise that may be eliminated if a post-process is added to suppress associations with function words.

4th experiment: The corpus studied here is a short biography of President *Barack Obama*. The text is written in Arabic. Upon reading the following table, it can be noted that in the text, *Obama* is strongly associated (M-confidence = 100%) to *Barack* even if the M-support is only 3. It is also noted that in terms of important values for M-confidence, *Obama* is strongly associated to the word pairs *origins, African* and *states, united*. However, there is a weak association of *Obama* with the function words *like* and *of* with an M-confidence of 66.67%. Once more, this type of noise can be eliminated with the addition of a post-process that would suppress the undesired associations.

TABLE IV.
RESULTS OF THE 4TH EXPERIMENT

Y	M-Support	M-Confiance
candidate, last	1	33,33 %
Arms	1	33,33 %
president life	1	33,33 %
Washington, American	1	33,33 %
Like	2	66,67 %
Of	2	66,67 %
states, united	2	66,67 %
origins, African	2	66,67 %
Barack	3	100,00 %

5th experiment: The experiments we will show (again we limit ourselves to show just maximal associations and scores of each association (M-support and M-Confidence). We assume that the reader is sufficiently familiar with the methods of classification and we do not need to show classes of similarities) were applied to a corpus extracted from Tripadvisor.fr. It consists of 45 guest reviews on a Parisian hotel. Reviews are written in French. The choice of this type of corpus is dictated by our desire to have a diverse vocabulary without being too large. The aim of these experiments is not to demonstrate the relevance of some classifiers or to demonstrate the ability of our approach to handle large documents but rather to assess the ability of our approach to extract strong associations between lexical units. To do this, several classifications have been produced using K-Means, Fuzzy-ART and SOM. The same vector representations were used as input for the three methods. To support a certain deformation data associated with the presence of errors in spelling we used tri-grams of characters as the unit of information. To reduce the number of tri-grams considered, we removed the tri-grams that contain spaces, numbers, special characters and those whose frequency of occurrence was less than 3. The size of the vector equals the number of distinct trigram enumerated in the corpus after cleaning. A vector was created for each corpus review. Moreover, to get the same number of classes (even if the classes, because of their vocabulary, are not similar) and have as much as possible the same basis of comparison, we set the classifiers in order to obtain 24 classes of similarities. For a sample of selected words (*Hôtel, Chambre, Déjeuner*), we tried to find in the classes of similarity obtained with all three classifiers, the words that co-occur with them. We find that despite the differences we observe in the results of the three classifications, the process of extracting the maximal association rules allows identifying the strongest associations which are spread over all classes obtained for each classification. In the case of our experiment, these associations appear to be, approximately, the same for all three classifications. We give in the tables below associations (whose m-support is greater than 1) extracted from classes obtained with the three classifiers. Higher are the values of M-Support and M-Confidence, stronger are the associations.

We note that in the case where X = *Hôtel*, the first 3 extracted associations are the same regardless of the

© 2013 ACADEMY PUBLISHER

classifier used (tables 5, 6 and 7). In the case where X = *Chambre*, 4 of the first 5 extracted associations are the same (tables 8, 9 and 10). In the case where X = *Déjeuner*, the first 2 extracted associations are the same regardless of the classifier used (tables 11, 12 and 13). However, if we just consider results of Fuzzy-Art and SOM, we note that 5 of the first 6 extracted associations are the same (tables 12 and 13).

TABLE V.
FIRST MAXIMAL ASSOCIATION RULES (CLASSES OBTAINED WITH K-MEANS)

X	Y	M-Support	M-Confidence
Hôtel	*Gare*	3	75%
	Chambre	3	75%
	Petit	3	75%
	Nuit	2	50%
	Dan	2	50%
	Nord	2	50%

TABLE VI.
FIRST MAXIMAL ASSOCIATION RULES (CLASSES OBTAINED WITH FUZZY-ART)

X	Y	M-Support	M-Confidence
Hôtel	*Petit*	6	100%
	Chambre	5	83,33%
	Gare	4	66,66%
	Personnel	3	50%
	Déjeuner	3	50%
	Salle	3	50%
	Nord	3	50%

TABLE VII.
FIRST MAXIMAL ASSOCIATION RULES (CLASSES OBTAINED WITH SOM)

X	Y	M-Support	M-Confidence
Hôtel	*Petit*	4	80%
	Chambre	4	80%
	Gare	4	80%

TABLE VIII.
FIRST MAXIMAL ASSOCIATION RULES (CLASSES OBTAINED WITH K-MEANS)

X	Y	M-Support	M-Confidence
Chambre	*Gare*	8	57,14%
	Petit	6	42,86%
	Nord	6	42,86%
	Salle	4	28,57%
	Hôtel	3	21,43%
	Dan	3	21,43%
	Déjeuner	3	21,43%

TABLE IX.
FIRST MAXIMAL ASSOCIATION RULES (CLASSES OBTAINED WITH FUZZY-ART)

X	Y	M-Support	M-Confidence
Chambre	*Petit*	8	58,82%
	Hôtel	6	29,41%
	Nord	6	35,29%
	Gare	4	41,18%
	Déjeuner	3	23,53%

TABLE X
FIRST MAXIMAL ASSOCIATION RULES (CLASSES OBTAINED WITH SOM)

X	Y	M-Support	M-Confidence
Chambre	*Gare*	10	62,5%
	Petit	9	56,5%
	Nord	7	43,75%
	Accueil	5	31,25%
	Hôtel	4	25%
	Déjeuner	4	25%
	Situer	4	25%

TABLE XI.
FIRST MAXIMAL ASSOCIATION RULES (CLASSES OBTAINED WITH K-MEANS)

X	Y	M-Support	M-Confidence
Déjeuner	*Petit*	3	100%
	Chambre	3	100%

TABLE XII.
FIRST MAXIMAL ASSOCIATION RULES (CLASSES OBTAINED WITH FUZZY-ART)

X	Y	M-Support	M-Confidence
Déjeuner	*Chambre*	4	100%
	Petit	4	100%
	Gare	3	75%
	Nord	3	75%
	Hôtel	3	75%
	Salle	2	50%
	Bain	2	50%

TABLE XIII.
FIRST MAXIMAL ASSOCIATION RULES (CLASSES OBTAINED WITH SOM)

X	Y	M-Support	M-Confidence
Déjeuner	*Chambre*	4	100%
	Petit	4	100%
	Gare	3	75%
	Hôtel	2	50%
	Salle	2	50%
	Situer	2	50%

Associations extracted from the classes of similarity are used to identify in the corpus relevant information that may represent the general labels of the object on which reviews were issued. To do this our system can identify in the corpus all segments that contain the associations and from which the labels will be extracted semi-automatically. In the case of our examples: *Hôtel* (hotel) is strongly associated with *Gare*, regardless of the classifier used, because the hotel is located next to a railway station (*gare*) and this association was often mentioned. It would become a label. It is important to mention that we show here that association rules approach can extract cooccurrences common to all three classifications. Therefore, these cooccurrences are called: strong associations. We completed our experiment with applying the process of association rules extraction over all classes obtained without distinguishing classifiers that allowed get them. Our initial goal was to verify the persistence of associations derived from classes in each classification. Our second goal was to test a process of meta-classification. We obtained the strong associations given in tables 14.

© 2013 ACADEMY PUBLISHER

TABLE XIV.
FIRST MAXIMAL ASSOCIATION RULES

X	Y	M-Support	M-Confidence
Hôtel	*Petit*	11	84,62%
	Chambre	10	76,92%
	Gare	9	69,23%
	Nord	6	46,15%
	Déjeuner	5	38,46%
	Accueil	4	30,76%
Chambre	*Petit*	22	53,66%
	Gare	21	51,22%
	Nord	16	39,02%
	Hôtel	10	24,39%
	Salle	8	19,51%
	Déjeuner	8	19,51%
Déjeuner	*Chambre*	8	100%
	Petit	8	100%
	Gare	6	75%
	Hôtel	5	62,5%
	Nord	5	62,5%
	Salle	4	50%

We note that the strongest associations are persistent. The "meta-classification" highlights common associations extracted from classes obtained by the three classifiers.

With these experiments, we have demonstrated that it is possible to extract strong associations in classes of similarity, regardless of the classifier used. These associations are relevant clues due to the regularity of their co-occurrence. The user (which is not necessarily an expert of the domain or a language engineer) can, according to the associations rules and their strength (given by the M-Support and the M-Confidence), select lexical descriptors that he/she thinks appropriate. Thus, in the examples discussed above, beside to *hôtel* as descriptor, the user can decide that *petit* and *gare* are important descriptors of the hotel. These descriptors remain clues that allow the user direct access to the part (or parts) of the text to which extracted associations refer.

All corpora we processed are heterogeneous. They are formed by contents of answers to an interview or web pages that meet a given query. The chosen strategy applies a text classification first then extracts maximal association rules (of the form $X \xrightarrow{\max} Y$) from the classes that contain query keywords. In this first step, X represents the query keywords. The user has several associations displayed. He chose one (or more) according to M-Support and m-Confidence scores but also to his own objective.

The process can be continued iteratively. Y becomes the new X for which maximal association rules will be identified.

At the end of the process the user can access the parts of the text that contain all the successive X.

We believe that association rules and maximal association rules employ measures that are generic enough and consistent to allow extraction of relevant associations hidden in noisy classes regardless of the classification method used. An association that frequently appears in classes generated with different classification methods (different classifiers or same classifier with different parameters) is called a strong association [24]. Such associations are useful to consolidate results obtained using different classification strategies. In sum, strong associations allow to highlight constant relations that can well describe the content and can be used as clues in computer assisted reading process.

V. CONCLUSION

Several strategies are employed to facilitate the exploration of textual documents. The labeling is one of the most common. The amount of textual documents available on networks required some mechanism to assist the selection of tags used. In this paper we had shown that association rules and maximal association rules can be applied to extract strong association in a set of classes of similarities. Strong associations can be considerate like stable descriptors of content. We believe that when the antecedent of a valuable rule is used as descriptor of a document it can be useful to add the consequent. Following that assumption can assist the selection of metadata. Because the proposed method is based on co-occurrence relations, it is not limited to providing assistance when labeling a document. The proposed method can be used for several other applications like lexical disambiguation, information retrieval, knowledge extraction and computer assisted reading.

In computer assisted reading, the user uses maximal association rules as clues in order to explore the text. He or she jumps from one part of the text to another considering the extracted associations. The reading here is assisted by the maximal association rules. The process remains under the control of the user.

ACKNOWLEDGMENT

This work was supported in part by a grant from Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] A. N. Srivastava and M. Sahami. *Text Mining, Classification Clustering and Applications*, Chapman & Hall/ CRC Press, 2009.
[2] J. Anderson. An Introduction to Neural Network, MIT Press, ISBN 0-262-01144-1, 1995.
[3] S. Haykin. Neural Networks: A Comprehensive Foundation, Macmillan College Publishing Company, ISBN 0-02-352761-7, 1994.
[4] A. Amir and Y. Aumann, *Maximal association rules: a tool for mining association in text*. Kluwer Academic Publishers, 2005.
[5] Y. Le Bras, P. Meyer, P. Lenca and S. Lallich. Mesure de la robustesse de règles d'association, In proceedings of the QDC 2010, Hammamet, Tunisie, 2010.
[6] B. Vaillant. Mesurer la qualité des règles d'association : études formelles et expérimentales, Thesis École Nationale Supérieure des Télécommunications of Bretagne, 2006.
[7] P. Hajek, I. Havel and M. Chytil. *The GUHA method of automatic hypotheses determination.* Computing, 1966.

© 2013 ACADEMY PUBLISHER

[8] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases", *In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, Proceedings of the 20th International Conference on Very Large Data Bases.* Santiago, Chile, 1994.
[9] R. Agrawal, T. Imielinski and A. Swami. Mining association rules between sets of items in large databases, *In Proceedings of the ACM SIGMOD International Conference on Management of Data*. Washington, 1993.
[10] S. Lallich and O. Teytaud. « Évaluation et validation de l'intérêt des règles d'association », *Revue des nouvelles Technologies de l'information,* 2003.
[11] H. Cherfi and A. Napoli. « Deux méthodologies de classification de règles d'association pour la fouille de textes », *Revue des nouvelles technologies de l'information*, 2005.
[12] C. T. Diop and M. Lo. « Intégration de règles d'association pour améliorer la recherche d'informations XML », *Actes de la Quatrième conférence francophone en Recherche d'Information et Applications.* Saint-Étienne, 2007.
[13] L. Rompré, I. Biskri and F. Meunier. "Text Classification: A Preferred Tool for Audio File Classification", *In Proceedings of the 6th ACS/IEEE International Conference on Computer Systems and Applications*, Doha, 2008.
[14] M. Damashek. "Gauging Similarity with n-Grams: Language-Independent Categorization Of Text", *Science*, 267, p. 843-848, 1995.
[15] G. Greffenstette. "Comparing Two Language Identification Schemes », *Actes des 3èmes Journées internationales d'Analyse statistique des Données Textuelles*, Rome, 1995.
[16] N. Patel and P. Mundur. "An N-gram based approach to finding the repeating patterns in musical", *In Proceedings of Euro/IMSA*, Grindelwald, 2005.
[17] W.K. Estes. Classification and Cognition, Oxford University Press, ISBN 0-19-510974-0, 1994.
[18] E. Miller, D. Shen, J. Liu, C. Nicholas and T. Chen. "Techniques for Gigabyte-Scale N-gram Based Information Retrieval on Personal Computers", In Proceedings of the PDPTA 99, Las Vegas, USA, 1999.
[19] G. Carpenter and S. Grossberg. Fuzzy ART: Fast Stable Learning and Categorisation of Analog Patterns by an Adaptative Resonance System. Neural Network, Volume 4, p. 759-771, 1991.
[20] J. MacQueen. "Some Methods for classification and Analysis of Multivariate Observations", In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1967.
[21] T. Kohonen. "The Self-Organisation Map", In Proceedings of the IEEE, Volume 78, No. 9, p. 1464-1480, 1990.
[22] N. Turenne. Apprentissage statistique pour l'extraction de concepts à partir de textes (Application au filtrage d'informations textuelles), Thèse de doctorat en informatique, Université Louis-Pasteur, Strasbourg, 2000.
[23] I. Biskri, L. Rompré, S. Descoteaux, A. Achouri and B. Amar Bensaber. "Extraction of Strong Associations in Classes of Similarities", In proceedings of IEEE/ICMLA, Boca Raton, Florida, USA, 2012.
[24] I. Biskri, H. Hilali and L. Rompré. "Extraction de relations d'association maximales dans les textes », In Proceedings of JADT 2010, p. 173-182 (2010)

Ismaïl Biskri: Full Professor at the department of Mathematics and Computer Science of the University of Quebec at Trois-Rivières (UQTR). He is researcher at LAMIA research group and Discourse and Communication research group. His interests in research concern Artificial Intelligence, Natural Language Processing, Combinatory Logics, Information Retrieval and Text-Mining.

Abdelghani Achouri: Ph.D. Student at the department of computer science of the University of Quebec at Montreal (UQAM). His interests are Information Retrieval and Text-Mining.

Louis Rompré: Ph.D. Student at the department of Computer Science of the University of Quebec at Montreal (UQAM). His interests are Information Retrieval and Data-Mining for music data.

Steve Descoteaux: Master Student at the department of Mathematics and Computer Science of the University of Quebec at Trois-Rivières (UQTR). His interests are Information Retrieval and Text-Mining.

Boucif Amar Bensaber: Full Professor at the department of Mathematics and Computer Science of the University of Quebec at Trois-Rivières (UQTR). He is researcher at LAMIA research group. His interests in research concern network.

© 2013 ACADEMY PUBLISHER