Using Strong Lexical Association Extraction in an Understanding of Managers' Decision Process

Frédéric Simard[1,2], Ismaïl Biskri[1], Josée St-Pierre[2], Boucif Amar Bensaber[1]

[1]LAboratoire de Mathématiques et Informatique Appliquées, Université du Québec à Trois-Rivières
[2]Canada Research Chair in risk and performance management of SMEs, Université du Québec à Trois-Rivières
C.P. 500, Trois-Rivières (QC) G9A 5H7, Canada
frederic.simard.10@ulaval.ca; {ismail.biskri; josee.st-pierre; boucif.amar.bensaber}@uqtr.ca

Abstract

Searching information or specific knowledge to understand decisions in a huge amount of data can be a difficult task. To support this task, classification is one of several used strategies. Algorithms used to support the process of automated classification leads to large and often noisy classes that is difficult to interpret. In this paper we present a method that exploits the notion of association rules and maximal association rules, in order to seek strong lexical associations in classes of similarities. We will show in experimentation section how these lexical associations can assist in understanding owner-managers decisions.

Introduction

When data is grouped into subsets, following an automatic or manual classification process, some regular patterns may appear. Studying these patterns can show unexpected properties of the data. These unexpected properties are hidden knowledge. Hidden knowledge is a source of information that can be used to bring out some tags properly describing the content of the data. Discovering hidden knowledge into classes of similarities is possible when an analysis is done by an expert. However, the number of classes, the noise contained in these classes and the often too large size of classes are barriers when analysis is performed on classification automatically generated. We suggest using the association rules and the maximal association rules to assist extraction of hidden knowledge, and a graphical tool (Gephi) to display this hidden knowledge.

We will experiment the obtained processing chain on transcripts of interviews of owner-managers of small businesses in order to assist the understanding of how they take their decisions. The following section reviews the definition of association rules and maximal association rules as defined respectively in Agrawal et al. (1993) and Biskri et al. (2010).

Maximal Association Rules

Association rules are used in data mining. They allow finding regularities in transactions. The interest for association rules mainly came from the work of Agrawal on transactional databases analysis (Agrawal et al., 1993). Agrawal has shown that it is possible to define rules to illustrate the relationship between items that co-occur in commercial transactions. He showed that association rules can be used to identify which products are frequently bought together. For instance clients who buy items *x* and *y* often buy item *z*. In that context the transactions are a list of products. However, the concept of transactions can be redefined in a more generalized form than the one suggested by Agrawal. In fact, a transaction can be simply defined as a finite subset of data, which allows applying association rules to many domains. The only challenge is to adapt the concept of transaction to targeted domain. When association rules are applied to classes of similarities then classes themselves are considered as transactions, or when association rules are applied to segments of one class of similarity then segments themselves are considered as transactions.

To explain what an association rule is, consider a set of data consisting of multiple transactions $T_i = \{t_1, t_2, t_3, \dots, t_n\}$. An association rule is denoted $t_i \rightarrow t_k$ where t_i is called the antecedent, t_k is called the consequent. It expresses relationship between t_i and t_k. The quality of an association is calculated using a measure m and a predefined threshold σ_m. Thus, a rule is considered as a good quality rule if $m(t_i \rightarrow t_k) \geq \sigma_m$. Several measures are proposed in the literature and many studies are dedicated to their evaluation (Lebras et al., 2010 ; Vaillant, 2006). Among existing measures, support and confidence are the most common. Support and confidence are defined as follow:

Support (X) : Let X be a subset of elements spread in different transactions, then the support of this subset is denoted $S(\mathrm{X})$. The support of X is equal to the number of transactions that contain X. The calculation of support is given by equation (1) where *n* equals the total number of transactions.

$$S(\mathrm{X}) = \sum_{i=1}^{\mathrm{n}} \delta up(T_i) \text{ where } \delta up(T_i) = \begin{cases} 0, & \mathrm{X} \not\subseteq T_i \\ 1, & \mathrm{X} \subseteq T_i \end{cases} \quad (1)$$

A subset X is considered frequent when $S(\mathrm{X})$ is greater than a predefined threshold σ_s.

Support $S(\mathrm{X} \rightarrow \mathrm{Y})$: If X and Y are two subsets of elements such that $\mathrm{X} \neq \emptyset$ and $\mathrm{Y} \neq \emptyset$ then the support of the association rule $\mathrm{X} \rightarrow \mathrm{Y}$ is denoted $S(\mathrm{X} \rightarrow \mathrm{Y})$. The support of an association rule is equal to the number of transactions that contain both X and Y. Thus, $S(\mathrm{X} \rightarrow \mathrm{Y}) \Leftrightarrow \mathrm{S}(\mathrm{X} \cup \mathrm{Y})$. The calculation of $S(\mathrm{X} \rightarrow \mathrm{Y})$ is given by equation (2) where *n* equals the total number of classes.

$$S(\mathrm{X} \rightarrow \mathrm{Y}) = \sum_{i=1}^{\mathrm{n}} \delta up(T_i) \text{ where } \delta up(T_i) = \begin{cases} 0, & \mathrm{X} \not\subseteq T_i \vee \mathrm{Y} \not\subseteq T_i \\ 1, & \mathrm{X} \subseteq T_i \wedge \mathrm{Y} \subseteq T_i \end{cases} \quad (2)$$

Confidence $C(\mathrm{X} \rightarrow \mathrm{Y})$: The confidence of an association rule $\mathrm{X} \rightarrow \mathrm{Y}$ is denoted $C(\mathrm{X} \rightarrow \mathrm{Y})$. The confidence of an association rule is equal to the number of transactions that contain both the antecedent and consequent among the transactions that contain the antecedent. The calculation of $C(\mathrm{X} \rightarrow \mathrm{Y})$ is given by equation (3).

$$C\left(\mathrm{X} \rightarrow \mathrm{Y}\right) = \frac{S\left(\mathrm{X} \rightarrow \mathrm{Y}\right)}{S(\mathrm{X})} \quad (3)$$

Typically association rules are extracted using APRIORI algorithm (Agrawal and Srikant, 1994). This algorithm used the support and confidence measures to restrict the number of associations considered. Association rules provide a significant advantage: they are able to extract hidden knowledge in classes of similarities or segments of one class of similarity even if they are numerous and noisy. Despite their potential, association rules may ignore relevant associations with low frequency of occurrence. For example, if an item X often appears with an item Y and less often with another item Z then it's probable that the association between X and Y is retained and not the association between X, Y and Z, the confidence of the relationship between X, Y and Z being too low compare with the relationship with X and Y.

A maximal association rules is denoted $\mathrm{X} \xrightarrow{max} \mathrm{Y}$. They are proposed to overcome the constraint related to the frequency of occurrence which applies to association rules. Maximal association rules are used to obtain exclusion associations. The form of this kind of association is given by equation (4) where S_1, S_2, S_3 and S_4 are subsets of data.

$$S_1 \cup \bar{S_2} \rightarrow S_3 \cup \bar{S_4} \quad (4)$$

Special measures are defined to establish the quality of maximal association rules (Amir et al., 2005). These measures are the M-Support and the M-Confidence.

M-Support $S_{max}(\mathrm{X})$: The maximal support or m-support of a subset of items X is denoted $S_{max}(\mathrm{X})$˙ If E_i and X are two subsets of elements such as $\mathrm{X} \subseteq E_i$, then $S_{max}(\mathrm{X})$ is equal to the number of transactions that contain X and no other item from Y.

$$E_i \cap T_i = \mathrm{X} \quad (5)$$

Consider the following elements to illustrate what is the m-support:

$T_1 = \{t_1, t_2\}$; $T_2 = \{t_3, t_5\}$; $T_3 = \{t_3, t_4, t_6\}$
$E_1 = \{t_1, t_2, t_3, t_4\}$
$\mathrm{X} = \{t_3\}$ where $\mathrm{X} \subseteq E_1$.

Given these elements, $S_{max}(\mathrm{X}) = 1$ because:

$E_1 \cap T_1 = \{t_1, t_2, t_3, t_4\} \cap \{t_1, t_2\} = \{t_1, t_2\} \neq \mathrm{X}$,
$E_1 \cap T_2 = \{t_1, t_2, t_3, t_4\} \cap \{t_3, t_5\} = \{t_3\} = \mathrm{X}$
$E_1 \cap T_3 = \{t_1, t_2, t_3, t_4\} \cap \{t_3, t_4, t_6\} = \{t_3, t_4\} \neq \mathrm{X}$.

M-Support $S_{max}\left(\mathrm{X} \xrightarrow{max} \mathrm{Y}\right)$: $S_{max}(\mathrm{X} \xrightarrow{max} \mathrm{Y})$ is the number of transactions that $S_{max}(\mathrm{X})$ and S(Y).

M-Confidence $C_{max}\left(\mathrm{X} \xrightarrow{max} \mathrm{Y}\right)$: The maximal confidence or m-confidence is given by equation (6) where |D(X, g(Y))| is a subset consisting of the transactions that $S_{max}(\mathrm{X})$ and that contains at least one item of Y.

$$C_{max}(\mathrm{X} \xrightarrow{max} \mathrm{Y}) = \frac{S_{max}\left(\mathrm{X} \xrightarrow{max} \mathrm{Y}\right)}{|\,\mathrm{D}(\mathrm{X},\ \mathrm{g}(\mathrm{Y}))|} \quad (6)$$

The algorithm used to extract maximal association rules is similar to the one used to extract regular association rules. It extracts rules whose m-support and m-confidence are above predetermined thresholds. Maximal association rules are complementary to the regular association rules. They highlight associations that regular rules tend to ignore. However, if only the maximal association rule is used, it may result in loss of interesting associations (Amir et al., 2005). The combined use of association rules and maximal association rules is very interesting during analysis of classes of similarities or during analysis of segments grouped in the same class because they are able to bring out various hidden knowledge.

We present our processing chain and methodology in the following section.

Methodology

Association rules and maximal association rules employ measures that are generic enough and consistent to allow extraction of relevant associations (often hidden in large and noisy classes) regardless of the classification method used. An association that frequently appears in classes generated with different classification methods (different classifiers or same classifiers with different parameters) is called a strong association. 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. The proposed methodology is to use the association rules and maximal association rules to extract recurring patterns within classes of similarities produced using various methods of classification. Captured associations are used like high level descriptors of the content. The process leading to the extraction of strong associations is mainly composed of the following six major steps: (1) Gathering of the data; (2) Preparation of the data; (3) Application of a classification method; (4) Extraction of strong associations; (5) Visualization of the results; (6) Evaluation of the results. Steps 2 and 3 can be repeated several times (with different settings) to create a large dataset. The three last steps are more related to the analysis of the data.

Gathering of the data: Prior to the experimentation is the gathering of the data used. In our particular study, we are considering a case of investment in Africa. Twenty owner-managers of small businesses with various profiles (experience, age, sex, domain, expertise) were invited to take part in our study by speaking their reasoning out loud on a project of business partnership implying two Cameroonian partners who exhibit different characteristics in terms of quality, profitability and competitiveness (firm A and firm B). The owner-managers were identified from the researchers' network and with the "snowball" strategy, where each owner had to refer another. The interviews were conducted at the workplace of each of the owners or at the Institute where they were isolated without phone nor computer, in presence of only the researcher.

The interviews were recorded, transcribed and then analysed with a prior objective of understanding how owner-managers take their decisions. The majority of owners took a final decision towards the project. Considering the analyses made upon the vocabulary of the interviews, some difficulties were encountered concerning the dissimilarity of the lexicons, reflecting variable knowledge of the French language among the owners.

Preparation of the data: The first step is to create vector representations of data. To do this, the text data is manually imported into the system. Imported data are segmented into sentences, paragraphs or freely depending of the configuration of the system. One vector representation is created per segment. The unit of information considered for building these vector representations is the word or the n-gram of characters. As a recall, n-gram of characters is defined as a sequence of *n* successive characters. For instance, in the word *computer* the 3-grams are: *com*, *omp*, *mpu*, *put*, *ute* and *ter*. The choice of n-gram of characters as a unit of information is justified by the fact that cutting into sequences of *n* consecutive characters is possible in most languages. Also n-grams of characters tolerate a certain ratio of distortion (Miller et al., 1999 ; Damashek, 1995). To restrict the size of the lexicon, some processing like deletion of hapax, lemmatization, deletion of stop words, etc. can be applied.

Application of a classification method: The second step is to apply a classification method among those available. Three classification methods are available: Fuzzy-ART, K-Means and SOM. In our research, the choice of these methods is not dictated by specific technical reasons. It is motivated by the interest of these methods in the literature (Anderson, 1995 ; Haykin, 1994). In fact, the modular architecture of our system encourages the integration of other methods of classification. After this step, classifications are produced. Classes gather together segments that are considered similar by the classifier used. They are associated with lists of words that represent the vocabulary of each class. This vocabulary is formed from the union (or intersection) of the vocabularies of the segments grouped in this class. Each class of similarity can be seen as a set of similar segments or as a lexicon formed from the lexicons of similar segments.

Extraction of strong associations: Extraction of strong associations is carried out using an adaptation of the method outlined in Biskri et al. (2010) where the similarity classes are treated as transactions. Here, the considered transactions are the segments grouped in each class. We are interested in extraction of strong associations in each class. The vocabulary of each class is used to build the subset E_1 in which X is selected. In other words, for each class E_1 is a subset of the vocabulary of this class, and $X \subseteq E_1$. In the simplest scenarios E_1 coincide with X. Otherwise, the value of X can be chosen by a user related to his objective (or randomly selected by the system).

Depending on X, the system generates the subset E_2. This subset is used to define the possible consequents. E_2 is built by subtracting X of the union of all transactions (segments grouped in the current class) where X appears. For illustration consider the following elements:

$T_1 = \{t_1, t_2, t_3,\}$ $T_2 = \{t_2, t_4, t_5\}$ $T_3 = \{t_1, t_5\}$

Say $X = \{t_1\}$, then $E_2^{'}$ equal the union of all transactions who contain t_1, more exactly the union of T_1 and T_3 or $\{t_1, t_2, t_3, t_5\}$. E_2 is created by subtracting X to $E_2^{'}$: $E_2 = E_2^{'} - X = \{t_2, t_3, t_5\}$. Each subset contained in E_2 is considered as a candidate. Thus, the subsets $\{t_2\}$, $\{t_3\}$, $\{t_5\}$, $\{t_2, t_3\}$, $\{t_2, t_5\}$, $\{t_3, t_5\}$ and $\{t_2, t_3, t_5\}$ may be consequents. Measures of support, confidence, m-support and m-confidence are finally calculated for each subset. To avoid a computational cost too significant, the cardinality of the consequents can be fixed.

Visualization of the results: Considering the large quantity of association rules that can be produced by the system, it can be essential to configure a visualization tool. The idea here is to represent each association as an edge on a directed graph.

Evaluation of results: Evaluation is the final step. The system provides an interface for easily navigating between rules, classes of similarities and their origin. The interface of the system is paired with the visualization tool for a better interpretation of the results.

Experimentations

To realize our experimentations we used a system previously developed to assist interpretation of classes of similarities (Biskri et al., 2010) and extended for a meta-classification process. This system was implemented in C#. The results of the analyses are stored in XML databases.

The main goal of this test case from a management point of view is to get a better understanding of the process of reflection owners-managers (hereafter OM) go through when facing a decision in a context of uncertainty. The idea of using a statistical system like ours is to be able to grasp the reasoning of a multitude of people.

Our experimentation was conceived after thoroughly reading all the transcriptions with the expectation that different investment choices were taken with different reasoning. Hence, certain patterns were noticed in the transcriptions of those who decided to invest with company A, as opposed to those with company B. These patterns made us consider the following experimentation: choose the parameters in order to regroup, after a classification process, the texts associated with the same decisions together in a single class and extract the association rules. To people who took the same decisions, who thought of the same concepts, association rules then represent relations between these concepts. The visualization tool helps to interpret these relations.

As stated previously, the goal of our experimentation is twofold: analysing the data as a helping tool for an understanding of OM's decisions process; studying and verifying the quality of our association rules extraction tool.

In order to attain these goals and considering the design of the software (which is conceived in order to access all words *word2* for a fixed word *word1* in associations $word1 \rightarrow word2$), we had to decide upon a set of words *word1* from which to evaluate the associations rules. This set of words was chosen as a subset of the full vocabulary present in the transcribed interviews. We asked the expert in finance to choose a subset of approximately ten words. The list (eleven words) is as follows (in French) : *compétitif, confiance, coopération, profit, projet, qualité, rentabilité, réputation, sécurité, stratégie, succès*. These loosely translate to: *competitive, trust, cooperation, profit, project, quality, profitability, reputation, security, strategy, success*. The preceding words were chosen with the hypothesis that they would have been used frequently enough to denote a statistically interesting idea made in the reasoning of the OM. We were looking for essential criterion to the OM for him to base his decision upon: be them profitability, security, quality, etc. The hypothesis appears to be sound; each and every of the eleven words are justified according to the context of the experiment. A preliminary interesting result is that only a few of them were used in a non-trivial association rule.

The lemmatizer available in the software was not used in the analyses because of the poor quality of some of the lemmatisations. Hence, all words were uniquely considered in the association rules, which bring up another limit: considering associations of the form $\mathrm{word1}_{\mathrm{singular}} \rightarrow \mathrm{word2}$ and $\mathrm{word1}_{\mathrm{plural}} \rightarrow \mathrm{word2}$ as different. Another unexpected aspect of our analyses was the frequent use of words such as "faq", which is a French slang for "then".

Our experiments were not mainly concerned with which classifiers were used, as it was showed in Biskri et al. (2013) that the three classifiers present in the software would lead to the same main association rules. This lead us to use one main classifier, ART, for the experiments. The main experiments for maximal association rules extraction were conducted with the first two classes obtained by the classifier ART, as they contained the most important transcripts. Transcripts of the first class were associated with the first decision whereas those of the second were associated with the second one.

At this step, obtained maximal rules are exported to Gephi tool via an excel file. Gephi is a free open-source software for graph visualization. It can take as input a table specifying the vertices of the graph and another its edges. The vertex table consists of a column of vertex names and another of vertex labels, if desired. The names in our case are the same as the labels. The edge table is made up of three columns, the first two specifying the end-points of the edge and the last its weight. The third column is optional, using it implies we needed to specify a measure for the weight of an edge. In our software, two measures are provided for each association, the m-support and the m-confidence. We quickly realized how poor the m-support is as a measure of quality of an association alone. We would sometimes find ourselves with m-support as low as 1 (absolutely, not relatively), meaning the association appeared only once in the whole text. However, we could not discard such associations so quickly, because another measure, the m-confidence, would in such cases frequently be up to 100%. We explained such results by considering the limited size of our sample (the transcripts) and the fact that some associations don't appear really often, but are still significant in their meaning.

This leads us to thinking a poor m-support is not necessarily a sign of a poor association rule. The reciprocal is also to consider: a poor m-confidence is not a good sign of an uninteresting association rule. In fact, m-support and m-confidence are considered here only as clues. In our case, since each segment represented a single transcript, we could consider, instead of an m-support, the number of segment in which an association is contained. This number is uniquely defined in our particular case. It can be regarded giving a slightly better idea than the m-support.

The following tables are examples of associations discovered in some classes. The first table comes from the transcripts taken from decision A, the second from decision B.

X	Y	M-Support	M-Confidence
marché	*accroitre*	4	100
	accroître	4	100
	améliorer	4	100
	associer	4	100
	cas	4	100
	concurrence	4	100
	élément	4	100
	endettement	4	100
	équipement	4	100
	euh	4	100
	facile	4	100
	intéresser	4	100
	fait	3	75
	façon	3	75
	marché	3	75
	procédé	3	75
	ssilence	3	75

Table 1 : *Association from transcripts with decision A*

X	Y	M-Support	M-Confidence
marché	*aller*	3	100
	ben	3	100
	chose	3	100
	dire	3	100
	faillir	3	100
	faire	3	100
	fait	3	100
	partir	3	100
	place	3	100
	produit	3	100
	qualité	3	100
	question	3	100
	savoir	3	100
	voir	3	100
	façon	3	100
	projet	3	100
	probablement	3	100
	terme	3	100

Table 2: *Association from transcripts with decision B*

The words in the tables show all those that were manually extracted from the graphical interface for a particular word X, here *marché*, French for *market*. In the graphical interface, words are presented in order of m-confidence from highest to lowest. Other words could have been extracted, only the 20 highest were noted.

As it appears in both tables, *marché* has an interesting out-degree; it leads with considerable m-confidence to many words. Some of them don't appear to carry particular meanings, for example *facile*, French for *easy*. There are however others, such as *équipement* (*equipment*), that form more interpretable associations. This particular association can be interpreted, in the case of decision A, as: OMs considering an investment with company A have to think about the necessity of modifying the company's process and their impact on the market after carrying such operation.

Making sense of all the data generated by the analyses of association rules can be hard considering the great amount of data and its sparseness. Here comes Gephi, the open-source graph visualization software. The matrix generated by the analyses of a specific decision (A or B) was initially considered as an adjacency matrix for constructing a graph to visualize our association rules. In order to use Gephi, we constructed vertex and edge tables, as described earlier.

Gephi offers a method for grouping the vertex of a graph in order to show a more interpretable (and visually appealing) structure. The details of these methods are not known and are of no concern here.

Below is the graph generated with the data from OMs having invested with company A.

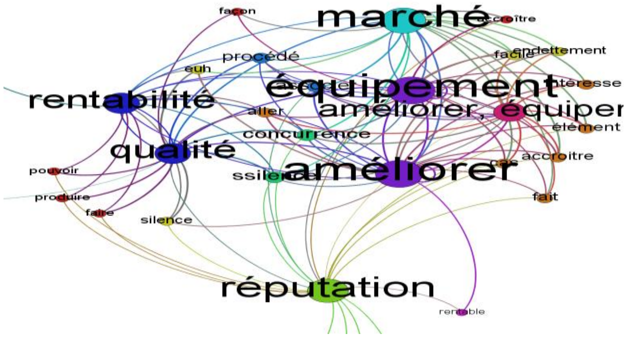

Figure 1: *Graph of data from owner-managers having invested with company A*

Some words were cut off from the graph in order to better show the important terms. As said previously, the size of vertex (or node) is proportional to its degree (the number of association rules it takes part in). Hence, few bigger nodes can then be considered as representing the important information OMs having invested with A expressed. The graph can be regarded as a compact view of association (edges) or of important words (vertices). Looking at the important words of the graph, we realize that a lot of them fall into common themes and are then unsurprisingly related with edges. Such words are, for example, *équipement, améliorer, qualité, rentabilité*. To anyone having read the case study and the answers given by the OMs, it seems trivial that these four words appear together: company A is making a great profit (“*rentabilité*”) with a product of poor quality (“*qualité*”), so the investor decides that the profit made by this company would be a great opportunity to upgrade (“*améliorer*”) its quality by acquiring better equipment (“*équipement*”).

The second graph, associated with decision B, shows five words that particularly stand out from the rest of the nodes, namely *marché, projet, qualité, rentabilité, réputation*. This is also an interesting result. Analysing the answers participants gave when investing with company B reveals a few patterns: people would consider the poor profitability (“*rentabilité*”) of the company not that important, since it is not alarming and because it has a reputation (“*réputation*”) of doing great product quality (“*qualité*”). They would also consider the market (“*marché*”) as being spread between company A and B, enhancing the quality of company B would enable them to gather better market-shares. Lastly, the word project (“*projet*”) appears often mostly because they are considering a project of investment, hence this word is basically part of their thought-process.

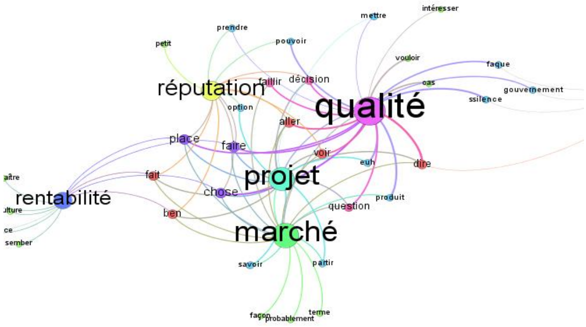

Figure 2 : *Graph of data from owner-managers having invested with company B*

On first regard, the second graph is clearly simpler than the first one; we might then conclude that the thought-process for investing with company B is also simpler than that for A. The transcripts reveal a totally different story. The investment towards company B is decided after considering many factors: this company makes great quality, has good equipment, the government is on its side, whereas company A produces poor quality, still A realizes more money than B, etc. Many variables seem to be considered before agreeing to invest in A or in B.

Conclusion

Several strategies are employed to facilitate the exploration of textual documents. The amount of textual documents available requires some mechanism to assist understanding of decisions process in order to eventually reapply them. In this paper we have shown that association rules and maximal association rules can be applied to extract strong lexical associations in a set of classes of similarities. Strong lexical associations can be considered like stable descriptors of textual document 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. As we have shown in this paper, following that assumption can improve the understanding of important decisions in business domain.

Acknowledgement

This work was supported by a grant from Natural Sciences and Engineering Research Council of Canada and by Canada Research Chair in risk and performance management of SMEs.

References

Agrawal, R.; Imielinski, T.; and Swami, A. 1993. Minning association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, 207-216.

Agrawal, A.; and Srikant, R.1994, Fast Algorithms for Mining Association Rules, *In Poceedings of the 20th International Conference on Very Large Database*, 487-499.

Amir, A.; Aumann, Y.; Feldman, R.; and Fresko, M. 2005. Maximal Association Rules: A Tool for Mining Associations in Text, *Journal of Intelligent Information System*, 333-345.

Anderson, J. 1995. *An Introduction to Neural Network*, MIT Press, ISBN 0-262-01144-1.

Biskri, I.; Hilali, H.; and Rompré, L. 2010. Extraction de relations d’association maximales dans les textes, *Actes du JADT*, 173-182.

Biskri, I.; Rompré, L.; Achouri, A.; Descoteaux, S. and Amar Bensaber, B. 2013. Seeking for High Level Lexical Association in *Texts*. *In Modeling Approaches and Algorithms for Advanced Computer Application* (eds) Amine A, Ait Mohamed O, Bellatreche L. Series : Studies in Computational Intelligence, Springer Publication.

Damashek, M. 1995. Gauging Similarity with n-Grams: Language-Independent Categorization of Text. *Science*, 267: 843-848.

Haykin, S.1994. Neural Networks: A Comprehensive Foundation, *Macmillan College Publishing Company*, ISBN 0-02-352761-7.

Le Bras, Y.; Meyer, P.; Lenca, P.; and Lallich, S. 2010. Mesure de la robustesse de règles d’association, *in Proceedings of the QDC 2010*, Hammamet, Tunisie.

Miller, E.; Shen, D.; Liu, J.; Nicholas, C.; and Chen, T. 1999. Techniques for Gigabyte-Scale N-gram Based Information Retrieval on Personal Computers, in *Proceedings of the PDPTA 99*, Las Vegas, U.S.A.

Vaillant, B. 2006. Mesurer la qualité des règles d’association : études formelles et expérimentales, Thesis École Nationale Supérieure des Télécommunications of Bretagne.