Extraction de relations d'association maximales dans les textes

Ismaïl Biskri, Hassane Hilali, Louis Rompré

Université du Québec à Trois-Rivières - Département de Mathématiques et Informatique

C.P. 500, Trois Rivières, Québec, Canada, G9A 5H7

Résumé

Dans cet article, nous présentons une approche dans laquelle nous combinons deux méthodes de fouille de données : la classification textuelle et les règles d'association maximales. La classification textuelle a longtemps été au centre de l'intérêt de plusieurs chercheurs. Toutefois, les résultats obtenus prennent la forme de listes de mots (classes) dont bien souvent on ne sait quoi faire. L'utilisation des règles d'association maximales présente plusieurs avantages dont : (i) la détection de dépendances et de corrélations utiles entre les unités d'informations (mots) des différentes classes ; (ii) l'extraction des connaissances cachées, souvent très pertinentes, à partir d'un grand volume de données.

Abstract

In this paper we will present a research on the combination of two methods of data mining: text classification and maximal association rules. Text classification has been for a long time the focus of interest of many researchers. However, the results take the form of lists of words (classes) that often we do not know what to do with. The use of maximal association rules induced number of advantages: (i) the detection of dependencies and correlations between the relevant units of information (words) of different classes, (ii) extraction of hidden knowledge, often relevant, from a large volume of data.

Keywords: classification, maximal association rules

1. Les règles d'association maximales

Un tour rapide de la littérature sur le data-mining (Amir and Aumann, 2005) nous enseigne que des règles d'association permettent de représenter les régularités de cooccurrences de données (au sens général du terme) dans des transactions peu importe leur nature. Ainsi, des données qui apparaissent régulièrement ensemble sont structurées dans des règles dites d'association. Une règle d'association est notée $X \Rightarrow Y$. Elle se lit comme suit : à chaque fois que la donnée X est rencontrée dans une transaction la donnée Y l'est aussi. Des mesures de qualité de ces règles d'association existent. Nous avons la mesure du *Support* et la mesure de la *Confiance.*

Les règles d'association représentent une notion encore récente. Des travaux sont menés pour juger de la pertinence des règles d'association ainsi que de la qualité de leur interprétation (Vaillant and Meyer, 2006; Lallich and Teytaud, 2003; Cherfi and Toussaint, 2002) ou encore de leur intégration dans des systèmes de recherche d'informations (Diop and Lo, 2007) ou dans des processus de classification pour la fouille de texte (Cherfi and Napoli, 2005).

Pour bien illustrer ce que sont les règles d'association, considérons la définition des principaux éléments à travers l'exemple suivant:

- Trois transactions pour regrouper les données qui co-occurrent : T1 : {A, 1, K} ; T2 : {M, L, 2} ; T3 : {A, 1, 2} ;
- Deux ensembles pour catégoriser les données : E1 : {A, M, K, L} ; E2 :{1, 2} ;
- X et Y deux ensembles disjoints d'unités d'information : X : {A} ; Y : {1}. $X \subseteq E1$ et $Y \subseteq E2$.

Pour une transaction Ti et un ensemble d'unités d'information X, on dit que Ti supporte X si $X \subseteq Ti$. Le Support de X, noté par S(X), représente le nombre de transactions Ti tel que $X \subseteq Ti$. Dans le cas des transactions T1, T2 et T3, S(X) = S(A) = 2.

Le *Support de la règle d'association* $X \Rightarrow Y$, est le nombre de transactions qui contiennent X et Y. Dans le cas de notre exemple $S(X \Rightarrow Y) = S(A \Rightarrow 1) = 2$.

La *Confiance de la règle d'association* $X \Rightarrow Y$, notée $C(X \Rightarrow Y)$ correspond au support de cette règle d'association divisé par le Support de X autrement dit $C(X \Rightarrow Y) = S(X \Rightarrow Y)/S(X)$. Dans le cas de notre exemple $C(X \Rightarrow Y) = C(A \Rightarrow 1) = 1$.

Malgré leur potentiel, les règles d'association ne peuvent être établies dans le cas d'associations moins fréquentes. Ainsi, certaines association sont ignorées car non fréquentes. Par exemple si le mot *imprimante* apparaît souvent avec le mot *papier* et moins souvent avec le mot *encre,* il est très probable que l'association entre *imprimante* et *papier* soit retenue au détriment de l'association entre *imprimante*, *papier* et *encre.* En effet, le critère de confiance associé à la relation entre *imprimante*, *papier* et *encre* serait trop bas.

Les règles d'association maximales que nous notons $X \xrightarrow{max} Y$ corrigent cette limite. Elles consacrent le principe général suivant : chaque fois que X apparaît seul, Y apparaît également. À noter que *X est réputé apparaître seul* si et seulement si pour une transaction Ti et un ensemble catégorie Ej *(X* $\subseteq$ Ej), $Ti \cap Ej = X$. Dans ce cas X est *maximale* dans Ti par rapport à *Ej* et Ti *M-Supporte* X. On note le *M-Support* de X par Smax(X), qui représente ainsi le nombre de transactions Ti qui M-Supporte X.

Dans la transaction T1, X n'est pas seul par rapport à E1 puisque $T1 \cap E1 = \{A, K\}$. Par contre, dans la transaction T3, X est seul puisque $T3 \cap E1 = \{A\}$.

Le *M-support* de l'association maximale $X \xrightarrow{max} Y$ notée par $Smax(X \xrightarrow{max} Y)$ représente le nombre de transactions qui M-supportent X et supportent Y.

Dans le cas de notre exemple seule la transaction T3 M-supporte X tandis que T1 et T3 supportent Y. Par conséquent $Smax\ (A \xrightarrow{max} 1) = 1$.

La *M-confiance* notée par $Cmax(X \xrightarrow{max} Y)$ représente le nombre de transactions qui M-supportent $X \xrightarrow{max} Y$ relativement à l'ensemble des transactions qui M-supportent $X \xrightarrow{max} E2$. La M-confiance de la règle $X \xrightarrow{max} Y$ est alors calculée par la formule $Cmax\ (X \xrightarrow{max} Y) = Smax(X \xrightarrow{max} Y)/\ Smax(X \xrightarrow{max} E2)$.

Dans l'association $A \xrightarrow{max} 1$, la M-Confiance se retrouve être égale à 0,5.

Enfin, il est à noter que nous devons définir des seuils minimums pour le M-support d'une association maximale ainsi que pour sa M-Confiance.

2. GRAMEXCO (les n-GRAms dans l'Extraction des Connaissances)

GRAMEXCO est un outil logiciel que nous avons développé pour la classification numérique des documents multimédia (Rompré et al., 2008) en particulier textuels. La classification numérique s'effectue au moyen d'un classifieur numérique. L'unité d'information considérée est le n-gram de caractères, la valeur de n étant paramétrable. L'objectif visé est de fournir la même chaîne de traitement, peu importe la langue du corpus, avec toutefois des aménagements dans la présentation des résultats pour en permettre une relative facilité de lecture. Pour rappel, l'utilisation des n-grams de caractères n'est pas récente. Elle prend naissance avec les travaux de Damashek (1995) sur l'analyse des textes et les travaux de Greffenstette (1995) pour l'identification de la langue. L'intérêt pour les n-grams aujourd'hui se retrouve étendu aux domaines de l'image (Laouamer et al., 2005) et de la musicologie en particulier dans le repérage des refrains (Patel and Mundur, 2005). On définira un n-gram de caractères par une suite de n caractères : bi-grams pour n=2, tri-grams pour n=3, quadri-grams pour n=4, etc. Par exemple pour le mot *informatique* les tri-grams sont : *inf, nfo, for, orm, rma, mat, ati, tiq, iqu, que*.

Le fonctionnement de GRAMEXCO n'est pas totalement automatique. Le choix de certains paramètres est fait par l'utilisateur en fonction de ses propres objectifs. GRAMEXCO prend en entrée un texte brut (non indexé) sous format UTF. Il s'en suit trois grandes étapes où l'utilisateur peut paramétrer certains traitements.

1. La **première étape** consiste à construire la liste des unités et des domaines d'information (segments de textes dont on veut comparer la similarité). Les deux opérations se faisant simultanément nous récupérons en sortie une matrice où sont répertoriées les fréquences d'apparition de chaque unité d'information dans chaque domaine d'information. Les unités d'information peuvent prendre la forme de bi-grams, de tri-grams, de quadri-grams, etc. L'obtention des domaines d'information passe par le processus de segmentation du texte qui peut se faire soit en mots, en phrases, en paragraphes ou tout simplement en sections de textes délimitées par un caractère ou une chaîne de caractères. Le choix de la taille du n-gram et du type du segment de texte revient à l'utilisateur en fonction de son objectif d'analyse.
2. La **deuxième étape** consiste à réduire la taille de la matrice. Cette opération est indispensable vue le coût important en termes de ressource que représenterait une matrice trop grande.

   Ainsi la liste des n-grams subit au cours de cette étape un nettoyage qui correspond à :

   - l'élimination des n-grams dont la fréquence est inférieure à un certain seuil ou supérieure à un autre seuil,
   - l'élimination de n-grams spécifiques sélectionnés dans la liste (par exemple des n-grams contenant des espaces ou des n-grams contenant des caractères non-alphabétiques),
   - l'élimination de certains n-grams considérés comme fonctionnels, par exemple les suffixes.
3. À la **troisième étape**, le processus de classification intervient. Le classifieur utilisé ici est le réseau de neurones ART (Meunier & al., 1997). Le choix de ce classifieur n'est pas dicté par des raisons de performances particulières car tel n'est pas notre objectif. Nous aurions tout aussi bien pu choisir un autre classifieur qui aurait certes donné des résultats différents. De telles variations continuent à faire l'objet de travaux de recherche comme ceux présentés dans Turenne (2000).

Au terme de cette étape des segments considérés comme similaires par le classifieur sont regroupés dans des classes de similarité. Aussi, le lexique de ces segments forme le vocabulaire des classes auxquelles ils appartiennent.

3. Identification de règles d'association maximales dans des classes de similarités

Que ce soit dans un objectif de désambiguïsation lexicale, ou de recherche de relations « conceptuelles », etc., l'interprétation des classes de similarité est un exercice non trivial (Fig. 1). Les classes de similarité sont généralement présentées comme des listes de mots qui co-occurrent ensemble. Ces listes sont bien souvent très volumineuses et malgré les aménagements apportés, leur vocabulaire reste très bruité.

Le processus d'extraction des règles d'association maximales s'avère des plus intéressants pour permettre la découverte d'associations lexicales pertinentes dans une prise de décision éclairée. Les classes obtenues au terme de l'opération de classification seront les transactions du processus qui nous permettront d'extraire les règles d'association maximales. Enfin, le processus, pour être mis en œuvre nécessite une supervision de l'utilisateur qui aura à déterminer en premier lieu le mot pour lequel il veut trouver les associations les plus vraisemblables.

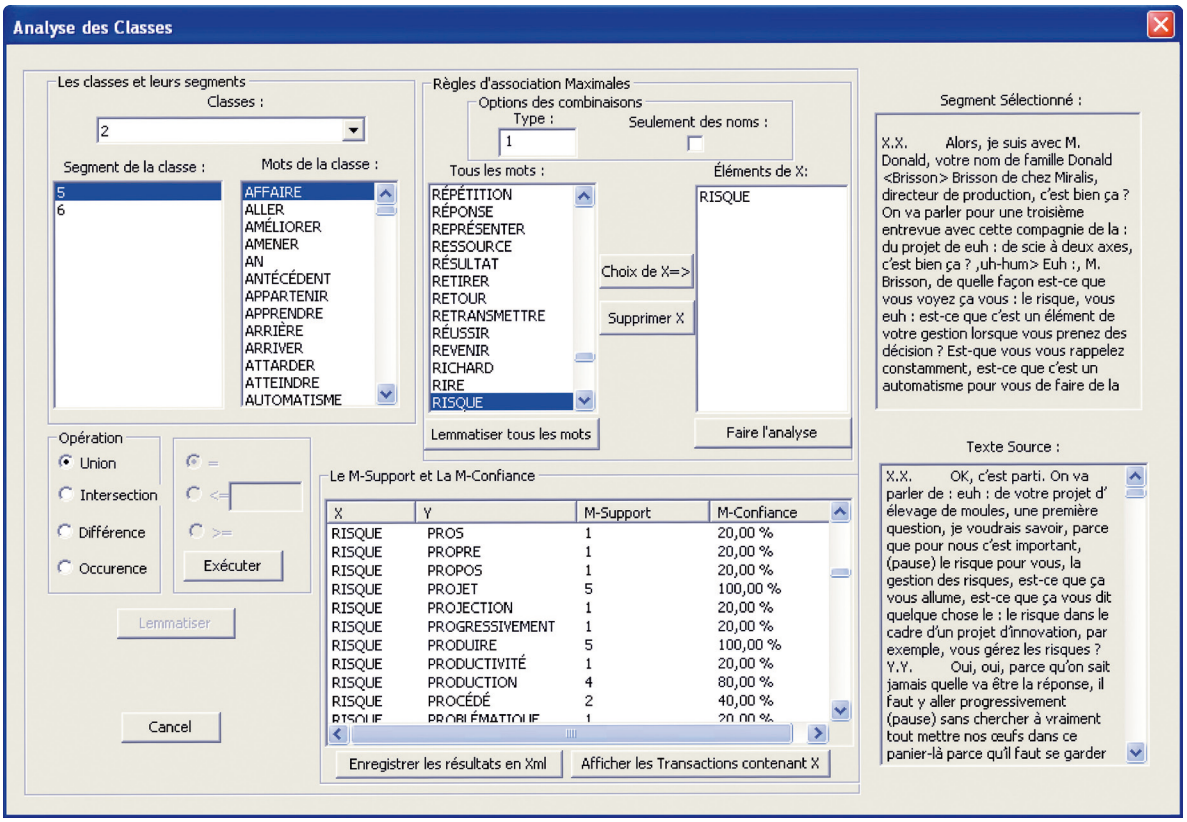

Figure 1 : Configuration du résultat d'une classification

Pour bien illustrer cette étape supposons le scénario suivant qui nous permettrait de découvrir des règles d'association maximales $X \xrightarrow{\max} Y$ à partir des résultats d'une classification.

Nous avons en entrée de la classification un texte dont le vocabulaire représente l'ensemble catégorie E1 : {x, a, b, c, d, e, f}. La classification donne en sortie les classes avec leur lexique respectif : C1 : {x, a, b, c}, C2 : {a, c, d}, C3 : {x, e, f, d}.

Si les classes représentent les transactions, le vocabulaire du texte en entrée représente un ensemble E1 pour catégoriser les données textuelles (le vocabulaire) dans lesquelles on choisit l'ensemble X.

Ceci étant établi, le processus d'extraction des règles d'association maximales s'effectue en trois étapes :

1ère étape : Choix de l'ensemble X : c'est l'utilisateur qui choisit parmi la liste des éléments de E1 le lexique qui va représenter notre X. Supposons pour les besoins de l'explication X = {x}.

2ème étape : Identification de l'ensemble Y et de l'ensemble E2 : l'identification de l'ensemble catégorie E2 dans lequel Y serait un sous-ensemble dépend fortement de l'ensemble X choisi et des classes dont X est un sous-ensemble.

Dans le cas de notre illustration X est inclus dans C1 et dans C3. Y peut alors être un sous-ensemble soit de {a, b, c} soit de {e, f, d}. Autrement dit, Y peut représenter un des sous-ensembles suivants : {a}, {b}, {c}, {a, b}, {a, c}, {b, c}, {a, b, c}, {e}, {f}, {d}, {e, f}, {e, d}. {f, d}, {e, f, d}.

Les mesures du M-Support et de la M-confiance seront calculées par rapport à ces différentes possibilités de valeurs de Y. Un processus itératif permettra de tester l'ensemble de ces possibilités. Nous pouvons, toutefois, limiter le nombre d'itérations pour éviter un coût computationnel trop prohibitif. Par exemple en fixant (au moyen d'un paramètre) la cardinalité du sous-ensemble Y.

Supposons que Y = {a, c}, pour construire E2 nous devons dans un premier temps établir les catégories respectives des éléments a et c. Celles-ci sont obtenues à travers l'union des classes qui contiennent a (respectivement c). Suite à quoi E2 = catégorie(Y) = catégorie {a, c} sera obtenu par l'intersection de la catégorie(a) avec la catégorie(c). Ainsi :

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

et

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

donc :

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

3ème étape : dès lors que les ensembles E1, E2, X et Y ainsi que les transactions ont été clairement identifiés, le calcul des mesures peut se faire.

Considérons l'association $x \xrightarrow{\max} a,c$. En utilisant les classes C1 : {x, a, b, c}, C2 : {a, c, d}, C3 : {x, e, f, d} comme transactions, et E2 = {a, b, c, d}, il en découle un M-support égal à 1, puisque seulement la classe 1 contient X= {x} et Y= {a, c}, et une M-confiance de 0.5 puisque deux classes contiennent X alors qu'une seule contient X et Y.

4. Expérimentations

Notre expérimentation a porté sur quatre corpus. Deux corpus sont en français et deux sont en arabe. Le premier corpus est un recueil d'entrevues avec des dirigeants de petites et moyennes entreprises québécoises pour connaitre leurs points de vue sur la notion du *risque*. Le deuxième corpus traite de l'histoire du règne du roi *Hassan 2*. Le troisième corpus (en arabe) traite de *l'organisation des pays exportateurs de pétrole*. Enfin, le quatrième et dernier corpus (en arabe)

résume la biographie du président américain *Barack Obama*. Les domaines sont suffisamment différents pour conclure sur la pertinence de la méthode.

1er expérimentation : le corpus comme mentionné précédemment porte sur le point de vue des dirigeants des petites et moyennes entreprises québécoises par rapport à la notion du *risque*. Une des contraintes au moment des entrevues est l'obligation faite aux dirigeants d'utiliser le mot *risque* là où ils le jugeraient nécessaire. Dans nos expérimentations, cet aspect est primordial car nous voulions savoir quels sont les mots associés au *risque* dans le discours des dirigeants.

Nous résumons les résultats obtenus dans le tableau suivant :

X	*Y*	*M-Support*	*M-Confiance*
Risque	Client	1	10%
	Actionnaires, Coût	1	10%
	Client, Projet	1	10%
	Décision, Produit	2	20%
	An	2	20%
	Marchés, Prix	2	20%
	Scie	3	30%
	Entrevue, Études	3	30%
	Fonction	4	40%
	Façon, Niveau	5	50%
	Produit	5	50%
	Question	6	60%
	Entrevue, Risque	6	60%
	Niveau, X	7	70%
	Gestion	7	70%
	Gestion, Projet	7	70%
	Projet, Risques	8	80%
	X	10	100%
	Pause	10	100%
	Projet, X	10	100%
	Pause, X	10	100%
	Projet	10	100%

Tableau 1 : Résultats de la 1ère expérimentation

Ainsi, malgré la présence de données bruitées, comme, par exemple, *Pause* et *X*, qui ont été insérées intentionnellement dans le texte pour des raisons d'éthique (*X* représente le nom des personnes questionnées) et pour représenter les silences (*Pause*), nous avons quand même des résultats très intéressants. Ainsi, par exemple

- *Risque* $\xrightarrow{\max}$ *Projet* est une association que nous retrouvons dans 10 classes (M-support = 10) avec une confiance de 100 %.
- *Risque* $\xrightarrow{\max}$ *Gestion*, *Projet* est une association que nous retrouvons dans 7 classes (M-support = 7) avec une confiance de 70 %. Autrement dit, il est possible à 30 % de trouver le mot *Risque* dans des classes où ne figureraient pas ensemble les mots *Gestion* et *Projet*.

- *Risque* $\xrightarrow{\max}$ *Gestion* est une association que nous retrouvons dans 7 classes (M-support = 7) avec une confiance de 70 %.
- *Risque* $\xrightarrow{\max}$ *Produit* est une association que nous retrouvons dans 5 classes (M-support = 5) avec une confiance de 50 %.

2ème expérimentation : Pour notre deuxième expérimentation nous avons choisi un court texte de 4 pages qui traite du règne du roi *Hassan 2*. Nous avons intentionnellement choisi pour cette expérimentation de considérer la cardinalité de l'ensemble Y égale à 1. Nous avons obtenu pour X = {*Hassan*} les résultats que nous résumons dans le tableau suivant :

X	*Y*	*M-Support*	*M-Confiance*
Hassan	Docteur	1	7.69 %
	Professeur	1	7.69 %
	Espagne	1	7.69 %
	Tunisie	1	7.69 %
	Espagnol	2	15.38 %
	Journaliste	3	23.08 %
	Histoire	3	23.08 %
	Préparer	3	23,08 %
	Titre	4	30,77 %
	France	5	38.46 %
	Politique	6	46.15 %
	Année	7	53,85 %
	Roi	8	61.54 %
	Maroc	8	61.54 %
	II	13	100 %

Tableau 2 : Résultats de la 2ème expérimentation

Nous constatons par exemple que l'association *Hassan* $\xrightarrow{\max}$ *II* est très forte. Sa confiance est de 100%. Il en est de même pour les associations *Hassan* $\xrightarrow{\max}$ *Maroc* et *Hassan* $\xrightarrow{\max}$ *Roi*. Bien que leur confiance n'est que de 61.54 %, celle-ci est suffisamment élevée pour considérer ces deux associations comme étant maximales.

3ème expérimentation : Pour notre troisième expérimentation nous avons choisi un texte en arabe qui traite de *l'organisation des pays exportateurs de pétrole* (OPEP). Notre souci étant d'évaluer la pertinence de la méthode par rapport à la langue arabe. Pour les besoins de l'expérimentation, nous avons choisi X = {«أوبك» (OPEP)}. Un résumé des résultats est donné dans le tableau 3.

Les résultats obtenus montrent bien l'étroite relation entre le mot acronyme أوبك (*OPEP*) et les deux mots منظمة (*Organisation*) et الدول (*Pays*). Toutefois, on constate une association avec un M-support et une M-confiance relativement élevés qui met en relation أوبك (*OPEP*) avec un mot fonctionnel في (*dans*). Nous considérons cette association comme étant un bruit, qui peut, toutefois, être éliminé si on rajoute un post-traitement qui supprimerait les associations avec des mots fonctionnels.

X	*Y*	*M-Support*	*M-Confiance*
OPEP أوبك	Mécanismes آليات	1	9,09 %
	Paris, Pays الدول باريس	1	9,09 %
	Création, tarifs أسعار إنشاء	2	18,18 %
	Pétrole البترول	3	27,27 %
	Pays, membres الأعضاء الدول	3	27,27 %
	Tarifs أسعار	3	27,27 %
	Organisation, tarifs أسعار منظمة	3	27,27 %
	Création إنشاء	3	27,27 %
	Membres الأعضاء	4	36,36 %
	Sommet قمة	4	36,36 %
	Monde العالم	4	36,36 %
	Organisation, pays الدول منظمة	4	36,36 %
	Organisation منظمة	6	54,55 %
	Pays الدول	7	63,64 %
	Dans في	9	81,82 %

Tableau 3 : Résultats de la 3ème expérimentation

4ème expérimentation : le corpus étudié ici est une courte biographie du président *Barack Obama*. Le texte est écrit en arabe. À la lecture du tableau ci-après nous constatons que *Obama* dans le texte est fortement associé (M-confiance = 100 %) à *Barack* même si le M-support n'est que de 3. Nous relevons également qu'en termes de valeurs importantes pour la M-Confiance *Obama* est fortement associé aux couples de mots (*origines, africaines*) et (*états*, *unis*). Toutefois, nous constatons des associations bruitées du mot *Obama* avec les mots fonctionnels comme et de avec une M-Confiance de l'ordre de 66,67 %. Encore une fois ce genre de bruit peut être éliminé par l'ajout d'un post-traitement qui supprimerait les associations non-désirées.

X	*Y*	*M-Support*	*M-Confiance*
Obama أوبـ اما	candidat, dernier آخر مر شح	1	33,33 %
	armes أ سـ لحة	1	33,33 %
	Vie président الـ رئـ يس حـ ياة	1	33,33 %
	Washington, américain أمريـ كي وا شـ نطن	1	33,33 %
	comme مـ ثل	2	66,67 %
	de من	2	66,67 %
	États, unis الـ مـ تحدة الـ ولايـ ات	2	66,67 %
	Origines, africaines أ صول أفـ ريـ قـ ية	2	66,67 %
	Barack بـ اراك	3	100,00 %

Tableau 4 : Résultats de la 4ème expérimentation

5. Conclusion

D'une façon générale les résultats de nos expérimentations semblent très recevables. L'utilisation en amont d'une classification numérique d'un processus d'extraction de règles d'association maximales peut aider à mieux lire les résultats d'une classification. La configuration de ces résultats semblait, en effet, décourager les chercheurs en sciences humaines qui se retrouvaient démunis en face de « listes volumineuses de mots ».

En outre, l'identification des associations maximales peut jouer un rôle majeur dans la résolution de problématiques auxquelles nous faisons face actuellement telle que la reformulation de requêtes, la construction semi-automatique et la maintenance d'ontologies, etc.

Enfin toute la théorie que nous avons présentée a été implémentée en C#. Les résultats des analyses sont stockés dans des bases de données XML. Aussi, nous envisageons à court terme de greffer un module de visualisation plus pratique qui permettrait de saisir sur un plan l'ensemble des associations d'une même unité lexicale.

Référence

Amir A. and Aumann, Y. (2005). *Maximal association rules: a tool for mining association in text*. Dordrecht: Kluwer Academic.

Cherfi, H. and Napoli A. (2005). Deux méthodologies de classification de règles d'association pour la fouille de textes. *Revue des nouvelles Technologies de l'Information.*

Cherfi H. and Toussaint Y. (2002). *Adéquation d'indices statistiques à l'interprétation de règles d'association. In JADT2002*, Saint-Malo.

Diop C.T. and Lo M. (2007). Intégration de règles d'association pour améliorer la recherche d'informations XML. In *Actes de la Quatrième conférence francophone en Recherche d'Information et Applications,* Saint-Étienne.

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

Greffenstette. G. (1995). Comparing Two Language Identification Schemes. In *JADT1995*, Rome.

Lallich S. and Teytaud O. (2003). Évaluation et validation de l'intérêt des règles d'association. *Revue des nouvelles Technologies de l'Information.*

Laouamer L., Biskri I. and Houmadi, B. (2005). Towards an Automatic Classification of Images : Approach by the N-Grams. In *Proceedings of WNSCI 2005*, Orlando.

Meunier J.G., Biskri I., Nault G. and Nyongwa, M. (1997). Exploration de classifieurs connexionnistes pour l'analyse terminologique. In *Actes de la conférence Recherche d'Informations Assistée par Ordinateur*, Montréal.

Patel N. and Mundur P. (2005). An N-gram based approach to finding the repeating patterns in musical. In *Proceedings of Euro/IMSA 2005*, Grindelwald.

Rompré L., Biskri I. and Meunier F. (2008). Text Classification: A Preferred Tool for Audio File Classification. In *Proceedings of the 6th ACS/IEEE International Conference on Computer Systems and Applications*, Doha.

Turenne N. (2000). *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, France.

Vaillant B. and Meyer P. (2006). Mesurer l'intérêt des règles d'association. *Revue des Nouvelles Technologies de l'Information (Extraction et gestion des connaissances: État et perspectives).*