CODAGE FLOU ET REPRESENTATION SYMBOLIQUE POUR L'ANALYSE D'UN CORPUS DE REPONSES A DES QUESTIONS OUVERTES

Sadika Rjiba[1] M. Gettler Summa [2] Saloua Benammou [3]

[1] *CML Faculté des Sciences Economiques et de Gestion Sousse - rjibasadika@yahoo.fr 1*
[2] *CEREMADE-CNRS Université Paris Dauphine - summa@ceremade.dauphine.fr 2*
[3] *CML Faculté des Sciences Economiques et de Gestion Sousse - Saloua.benammou@yahoo.fr 3*

Résumé. Après une démarche classique de prétraitement de données textuelles, nous recourons d'une part au Codage Flou et d'autre part à une représentation des données dans le cadre de l'Analyse Symbolique des Données pour analyser un corpus de réponses à des questions ouvertes. Nous définissons un univers de discours pour chaque mot clé retenu, en fonction des formes graphiques constituant notre corpus, chacune de ces formes graphiques étant caractérisée par un degré d'appartenance. Nous proposons en particulier des définitions de la notion de 'contexte d'un mot' sous les deux approches, floue et symbolique, ainsi que la construction des tableaux qui s'en déduisent afin d'analyser l'information contenue dans le corpus textuel. Nous illustrons cette méthode sur un questionnaire comportant des réponses à des questions ouvertes administrées auprès des universitaires tunisiens. Ce questionnaire concerne le ressenti de la révolution tunisienne de 2011.

Mots-clés. Analyse des Données Textuelles, Analyse Symbolique des Données, Codage flou, questions ouvertes

Abstract. After a conventional textual data preprocessing, we use on one hand fuzzy coding and also a representation of data in the framework of the Symbolic Data Analysis, to analyze a corpus of answers to open ended questions. We define a universe of discourse for each chosen keyword, based on the graphical forms that constitute the corpus; each of these graphic forms are characterized by a degree of membership. We propose in particular formalizations for the concepts of 'context of a word' under both approaches, fuzzy and symbolic, as well as the construction of tables that are deduced to analyze the information contained in the text corpus. We illustrate this method on a questionnaire with answers to open-ended questions administered from Tunisian universities. This questionnaire concerns the academic staff feelings about the Tunisian revolution of 2011.

Keywords. Textual data, Symbolic Data Analysis, Fuzzy Logic Coding, open-ended questioning

1 Introduction

Le traitement statistique d'un texte est une problématique étudiée depuis longtemps (Muller, 1968). Dans ce contexte, le débat reste ouvert sur la lemmatisation (Salton, 1989) des données, dans la phase de prétraitement. Cette technique (Labbé, 2001) désigne l'analyse lexicale d'un corpus facilitant le regroupant les mots d'une même famille. Chacun des mots d'un contenu se trouve ainsi réduit en une entité appelée lemme. Dans cette étude nous nous limitons à analyser le corpus dans son état brut (sans lemmatisation). Le listage des formes graphiques qui apparaissent dans le texte à analyser est essentiel il permet de calculer les fréquences d'apparition, ce qui facilite la présentation des segments répétés (Lebart et Salem, 1988) selon les approches déjà courantes dans la communauté de l'Analyse des Données Textuelles (Benzécri & al.1981).

La théorie des ensembles flous (Zadeh, 1965) est une approche qui a enrichi l'analyse décisionnelle, la représentation des connaissances et la classification des formes, ainsi que les systèmes experts.

Nous faisons recours à cette approche afin de définir un univers de discours pour chaque mot clé sélectionné, en fonction des formes graphiques constituant le corpus, chacune de ces formes graphiques étant caractérisée par un degré d'appartenance.

On cherche alors également une représentation dans le cadre de l'Analyse des Données Symbolique (ADS) (Billard et Diday, 2006), de la connaissance que constituent les contextes des mots clés. L'ADS est adaptée au traitement des données complexes, et dans ce cadre les individus statistiques répondent aux diverses définitions possibles pour les objets symboliques (Diday, 1995).

2 Définitions contextuelles des occurrences

On appelle **occurrence-phrase** d'un mot du corpus relativement à un mot clé, le nombre de fois ou un mot est associé au mot clé à l'intérieur d'une même phrase.
On appelle **occurrence k-mots** d'un mot du corpus relativement à un mot clé, le nombre de fois ou un mot est associé au mot clé à k-mots avant et à k-mots après le mot clé.
On appelle **occurrence k-lettres** d'un mot du corpus relativement à un mot clé, le nombre de fois ou un mot est associé au mot clé à k-lettres avant et à k-lettres après le mot clé.
Nous construisons un tableau de contingence des mots clés $\{M_i, i\in\{1,\ldots p\}\}$ avec les N mots du corpus selon l'une des trois définitions. Les co-occurrences aussi, sont calculés selon l'une des trois définitions contextuelles.

$\{V_1, V_i, V_k\}$ sont des sous ensembles des mots clés, sélectionnés a priori de façon arbitraire et a posteriori dans les applications, à partir des mots ordonnés par occurrences décroissantes dans le corpus et regroupés par des sémantiques pertinentes pour la problématique de l'étude.
Par exemple dans l'enquête que nous présentons dans le dernier paragraphe en application et qui concerne la révolution tunisienne, ses causes déclencheuses et ses conséquences (économiques, politiques, etc.), nous avons sélectionné les mots *'révolution'* et *'tunisienne'* comme étant les mots clés de cette analyse

- parce que d'une part ils appartiennent aux vingt mots de fréquences les plus élevées.
- et parce que d'autre part ils sont au cœur du sujet de l'étude

Par analogie à l'analyse des correspondances multiples, nous obtenons le tableau de contingence ci-dessous où tous les mots sont en lignes et les 'contextes' dépendant des mots clés en colonnes.

TABLEAU 1 : Tableau d'entrée en ADS pour des données textuelles

	V_1	V_2	…	V_j	…	V_k
	$M^c_{1,1} \ldots M^c_{p_1,1}$	$M^c_{1,2} \ldots M^c_{p_2,2}$	…	$M^c_{1,j} \ldots M^c_{p_j,j}$	…	$M^c_{1,k_j} \ldots M^c_{p_j,k_j}$
m_1	$O_{1,p_1,1}$	$O_{1,p_1,2}$	…	$O_{1,p_1,j}$	…	O_{1,p_1,k_j}
	⋮	⋮	…	⋮	…	⋮
m_i	$O_{i,p_1,1}$	$O_{i,p_2,2}$	…	$O_{i,p_j,j}$	…	O_{i,p_j,k_j}
⋮	⋮	⋮	…	⋮	…	⋮
m_{N-p}	$O_{N-p,p_1,1}$	$O_{N-p,p_2,2}$	…	$O_{N-p,p_j,j}$	…	O_{N-p,p_j,k_j}
M^c_1	$O_{M_1,p_1,1}$	$O_{M_1,p_1,2}$	…	$O_{M_1,p_1,j}$	…	O_{M_1,p_1,k_j}
⋮	⋮	⋮	…	⋮	…	⋮
M^c_p	$O_{M_p,p_1,1}$	$O_{M_p,p_1,2}$	…	$O_{M_p,p_1,j}$	…	O_{M_p,p_1,k_j}

Ce tableau de contingence (chaque cellule porte une fréquence) nous permet de déduire des formalisations lignes, dont l'expression est dans le cadre de l'ADS et des formalisations colonnes

dont les valeurs correspondent à un codage flou. L'intérêt de ces représentations est d'utiliser dans un deuxième temps les approches de statistique exploratoire et d'aide à la décision spécifiques de l'ADS et de la logique floue qui sont moins réductrices de l'information qu'un codage monovalué.

2.1 Les lignes : Approche symbolique

Les profils lignes (tableau 2) sont les contextes de m_i selon les sous ensembles des mots clés. Ainsi, la marge est le nombre total des citations du mot m_i dans les définitions contextuelles de l'ensemble des mots clés. Une description d'un mot appartenant au contexte étudié s'écrit sous la forme suivante :

$$d(m_i) = \left\{ \left[V_1 = \left\{ \cup M_{j,1,o_{i,j,1}} \right\}_{j=1} \right] \cap \left[V_2 = \left\{ \cup M_{j,2,o_{i,j,1}} \right\}_{j=2} \right] \cap \ldots \left[V_i = \left\{ \cup M_{j,j,o_{i,j,j}} \right\}_{j=j} \right] \cap \ldots \left[V_k = \left\{ \cup M_{j,k,o_{i,j,k}} \right\}_{j=k} \right] \right\}$$

Avec :

$d(m_i)$: La description d'un mot de contexte en fonction des mots clés sélectionnés.
$\{V_1, V_i, V_k\}$: Les sous ensembles des mots clés sélectionnés sont constitués a priori de façon arbitraire et a posteriori dans les applications en ordonnant les occurrences des mots.
$\{\cup, \cap\}$: Opérateurs logiques d'union et d'intersection
$\left\{M_{j,1,o_{i,j,1}}, \ldots, M_{j,j,o_{i,j,j}}, \ldots, M_{j,k,o_{i,j,k}}\right\}$: les mots clés des différents segments.

TABLEAU 2 Récapitulatif ADS

<table>
<tr><td></td><td>Analyse symbolique sur DT</td></tr>
<tr><td>variables</td><td>Mots clés</td></tr>
<tr><td>Modalités</td><td>fréquences</td></tr>
<tr><td>équations</td><td>$M_C^p \left\{ m_1 f_p^{o_1}, m_2 f_p^{o_2}, \ldots, m_i f_p^{o_i}, m_p f_p^{o_p} \right\}$</td></tr>
<tr><td>Tableaux des données</td><td>
<table>
<tr><td></td><td>M_C^1</td><td>M_C^i</td><td>M_C^p</td></tr>
<tr><td>m_1</td><td>$f_1^{o_1}$</td><td>$f_i^{o_1}$</td><td>$f_p^{o_1}$</td></tr>
<tr><td>m_2</td><td>$f_1^{o_2}$</td><td>$f_i^{o_2}$</td><td>$f_p^{o_2}$</td></tr>
<tr><td>m_i</td><td>$f_1^{o_i}$</td><td>$f_i^{o_i}$</td><td>$f_p^{o_i}$</td></tr>
<tr><td>m_p</td><td>$f_1^{o_p}$</td><td>$f_i^{o_p}$</td><td>$f_p^{o_p}$</td></tr>
</table>
</td></tr>
</table>

2.2 Les colonnes : Approche floue

La lecture colonne comporte le codage flou (Moreau & al. 2000, Goldfarb & al. 2001) de chaque mot clé sur le corpus, tout ou partie. C'est l'univers du discours de chaque mot clé (tableau 3).

TABLEAU 3 Récaptulatif *LF*

	Logique floue sur DT
variables	Ensembles flous
Modalités	Degré d'appartenance
équations	$M_C^p \{ m_1 d_p^{o_1}, m_2 d_p^{o_2}, \dots, m_i d_p^{o_i}, m_p d_p^{o_p} \}$
Tableaux des données	M_C^1 | M_C^i | M_C^p m_1 : $d_1^{o_1}$ | $d_i^{o_1}$ | $d_p^{o_1}$ m_2 : $d_1^{o_2}$ | $d_i^{o_2}$ | $d_p^{o_2}$ m_i : $d_1^{o_i}$ | $d_i^{o_i}$ | $d_p^{o_i}$ m_p : $d_1^{o_p}$ | $d_i^{o_p}$ | $d_p^{o_p}$

3 Application

Dans ce travail nous présentons des résultats de l'analyse des données textuelles appliquée à une problématique de Sciences Economiques induite par l'évènement 'révolution tunisienne' en 2011.
Les données ont été collectées lors d'une enquête par questionnaire auprès d'un échantillon de taille 541. Il s'agit d'étudiants ou de jeunes professionnels: hommes et femmes d'origine très diverses, mais tous inscrits aux facultés (institut, école, etc.) des universités de Sousse et de Monastir. Le terrain ne respecte pas exactement les quotas signalétiques, en particulier en ce qui concerne les divers lieux d'étude proposés au Sahel, le tirage des répondants n'est pas non plus fait au hasard stricto sensu. Cependant, même si nous n'avons pas remarqué de biais avéré dans l'échantillon, nous utiliserons principalement les données dans un but d'application des méthodes d'analyse textuelles quantitatives proposées, et non pas pour les conclusions politico-sociologiques éventuelles de l'enquête. Cette enquête à été réalisée au cours de l'année universitaire 2012/2013, par interviews directes en face à face. La durée d'une interview est en moyenne d'une heure.

Quatre questions ouvertes sont posées, les deux premières sont sous forme de commentaires se rapportant à deux questions fermées codées par échelle de Lickert, en particulier relativement au niveau de fierté du répondant sur son vécu de la révolution et sur le fait d'être un citoyen tunisien. La troisième question ouverte explicite la question fermée classant les causes majeures de déclanchement de la révolution tunisienne. Finalement, une dernière question ouverte concerne l'avis du répondant sur la situation économique du pays après la révolution.

TABLEAU 4 Exemple de tableau de données combinant *LF* et *ADS*

	$M_1^c = Révolution$	$M_2^c = Tunisien$
$m_1 = Chômage$	$d_{11} = 0.20$	$d_{12} = 0.30$
$m_2 = Inégalité$	$d_{21} = 0.10$	$d_{22} = 0.25$
$m_3 = Dictature$	$d_{31} = 0.20$	$d_{32} = 0.25$
$m_4 = dignité$	$d_{41} = 0.50$	$d_{42} = 0.20$

Nous représentons ici en exemple l'ensemble flou $Révolution$ (lecture en colonne). Les degrés d'appartenances de chaque élément de l'ensemble flou sont $\{d_{11}, d_{21}, d_{31}, d_{41}\}$. Ainsi l'ensemble flou *Révolution* est de la forme suivante :

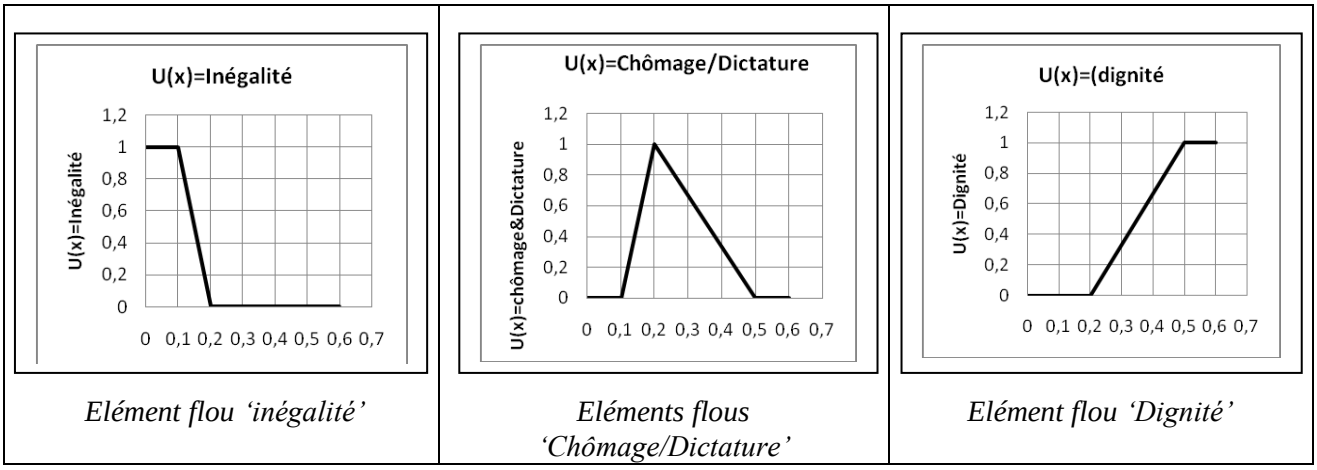
$$Révolution = \{Chômage(0.2); Inégalité(0.1); Dictature(0.2); Dignité(0.5)\}$$

Une lecture ligne du tableau n° 4, permet autant de représenter symboliquement les éléments de l'ensemble flous *'Révolution'* (m_1, m_2, m_3, m_4) en fonction des mots clés (M_1^c, M_2^c). Ainsi nous obtenons les équations suivantes :

$chômage = \{M_1^c 0.2, M_2^c 0.3\}$; $inégalité = \{M_1^c 0.1, M_2^c 0.25\}$; $dictature = \{M_1^c 0.2, M_2^c 0.25\}$;

$diginité = \{M_1^c 0.5, M_2^c 0.2\}$. Nous illustrons dans le tableau ci-dessous les différents éléments de l'ensemble flou *'révolution'*. Ces représentations graphiques, valeurs en abscisses et en ordonnées, sont effectuées par analogie aux représentations graphiques des ensembles flous en logique floue (Gogévac 1999). On ordonne ici les éléments par fréquences croissantes en relation au mot clé 'révolution' L'union des trois représentations floues de l'univers *'révolution'* est affichée dans la figure n°1, et l'on obtient l'ensemble flou 'révolution' dans le contexte de l'enquête de l'application.

TABLEAU 5 Représentation flou des éléments de l'univers de discours *'Révolution'*



Elément flou 'inégalité' | *Eléments flous 'Chômage/Dictature'* | *Elément flou 'Dignité'*

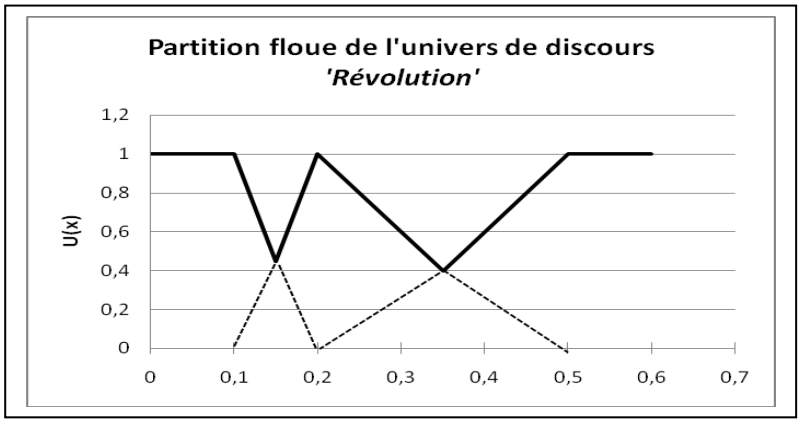


Figure 1 : *Ensemble flou 'Révolution'*

Un algorithme est en cours de test pour construire les diverses représentations et les tableaux qui s'en déduisent afin de procéder ensuite à l'analyse textuelle symbolique d'une part et floue d'autre part des données.

Bibliographie

[1] Diday, E. (1995) Probabilistic Objects for a Symbolic Data Analysis, Series in Discrete Mathematics and Theoretical Computers, 19.

[2] Billard, L. et Diday, E. (2006), Symbolic Data Analysis, Wiley.

[3] Lukasiewicz, J. (1920), English translation: On three-valued logic, *in L. Borkowski (ed.), Selected works by Jan Lukasiewicz*, North–Holland, Amsterdam.

[4] Knuth, D. (1998), Art of Computer Programming, *Volumes 1–3 Boxed Set (2nd Edition)*, Addison Wesley Professional.

[5] Black, M. (1937), Vagueness, An Exercise in Logical Analysis . *Philosophy of Science*, 427–455.

[6] Gojevac, J. (1999), Idées nettes sur la logique floue, 9-18.

[7] Meunier, B. B. (1995), La logique floue et ses applications, Addison-Wesley.

[8] Zadeh, L. (1994), Fuzzy logic, Neural Networks and soft computing, *ACM*, Vol 37, n°.3.

[9] Esnault, J. M. , Vermandel, M., Morchhauser, F., Steinling, M. et Huglo, D. (2007), Détermination par logique floue des volumes tumoraux en TEP : Application au suivi de la radio-immunothérapie des lymphomes. *Science direct, medecine nucléaire*. 656-664.

[10] Lebart, L. et Salem, A. (1994), Statistiques textuelles, Dunod.

[11] Reinert, M. (1993), Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars », *Langage et Société*, n°66, 5-39.

[12] Ahanda, B. T. (1998), Extension des méthodes d'analyse factorielle sur des données symboliques *Thèse de doctorat*, Université Paris Dauphine.

[13] Salton, G. (1989), Automatic text processing, Addison Wesley, USA.

[14] Zadeh, L.A. (1965), Fuzzy sets, *information and control*, 8, 338-353.

[15] Muller, C. (1968), Initiation à la statistique linguistique, Larousse, Paris.

[16] Goldfarb, B. et Pardoux, C. (2001), Étude de données multidimensionnelles évolutives et comparaison de codages par l'analyse factorielle multiple, *Revue de Statistiques Appliquée*, Vol 49, 97-117.

[17] Moreau, J., Doudin, P. A. et Cazes, P. (2000), l'Analyse des Correspondances et les Techniques Connexes, Springer.

[18] Benzécri, J. P. et Collaborateurs (1981), Pratique de l'Analyse des Données, Linguistique et Lexicologie, Dunod.

[22] Labbé , D (2001), Normalisation et lemmatisation d'une question ouverte. Les femmes face au changement familial. *Journal de la Société Française de statistique*, 37- 58.