ÉTUDE EXPÉRIMENTALE COMPARATIVE DES MÉTHODES STATISTIQUES POUR LA CLASSIFICATION DES DONNÉES TEXTUELLES

Ilham Benhadid, Jean-Guy Meunier, Saâd Hamidi, Zira Remaki, Moses Nyongwa
Laboratoire de l'ANalyse Cognitive de l'Information
Université du Québec à Montréal
Meunier.jean-guy@uqam.ca
{Ilham, saad, zira, nyongwa }@pluton.lanci.uqam.ca

Publié dans . JADT Nice 1998.

Abstract : The main task of this paper is to propose a comparison between classifiers (K-means, Markov, Art) on the textual data. Indeed, the goal is to provide an assistance to attain some aspects of the informational or semiotic content of a text (discursive, lexical, hypertextual, etc). Our document is composed from a merge between a spirale text (Belgium) and Hydro-Québec text. The methodology consists in to extract the related Hydro-Québec text from spirale text. We should mention that these two text do not share the same semantics. The results indicated that there is no big major difference between all the applied classifiers. For this reason, in the future perspective works the emphasis will be put on the computational and algorithm simplicity in order to improve the processing time.

1. Introduction

De nos jours, l'organisation du contenu de gros corpus en des configurations interprétables s'impose en vue de faciliter et d'assister l'analyste (terminologue ou autre) dans ses tâches de navigation et d'exploration. Dans notre perspective, l'extraction des connaissances peut être vue comme un processus de traitement classificatoire qui identifie des segments de textes contenant un ″même″ type d'information. Les tâches de dépistage, d'exploration et de récupération de l'information ou "connaissances" contenues dans ces textes sont devenues extrêmement ardues et de moins en moins possibles dans des temps raisonnables. Or, le problème qui retient notre attention dans l'analyse de textes par ordinateur est précisément celui du processus par lequel on peut, dans un texte, accéder à son contenu original et en extraire ces dites connaissances. Dans cet article, nous proposons une étude comparative de certains classifieurs dont le but commun est de créer des classes de segments de textes de contenu sémantique similaire.

2. Problématique

La littérature technique tant de la gestion documentaire (Salton 1994) que de l'analyse du contenu (Virbel 1993, Delany 1993 Lenhert 1994, Croft 1995, Bookman 1994) montre qu'il est devenu impératif de développer des outils permettant d'atteindre non seulement le document ou le segment d'un texte mais leur contenu. Certaines méthodes sont symboliques, i.e. procèdent par règles et/ou par grammaire (Sabbah 1991, Lenhert, 1995, Croft 1994, Desclées et 1995 et col.). D'autres sont statistiques (Church 1991, Wilks 1994, Veronis, 1991, Balpe et Lelu, 1995, Salem et Lebart 1993 etc). Un des problèmes majeurs cependant de ces méthodes est celui de pouvoir effectuer cet accès en un temps raisonnable sans toutefois altérer leur pouvoir classificatoire et ceci, même sur de gros corpus.

3. Matériel et Méthode

L'étude a porté, d'une part sur 56 pages de la revue spirale (Belgique) dans lesquelles nous avons introduit 4 pages du texte d'Hydro-Québec (texte A), et d'autre part sur un plus gros corpus constitué des 180 pages de spirale et de 40 pages d'Hydro-Québec (texte B). Nous avons procédé

ensuite à une transformation des textes pour en obtenir une représentation matricielle. A chaque segment du texte (32 lignes) correspond une ligne de la matrice dont chaque élément indique, pour un lemme donné, la fréquence de son apparition dans le segment. Une méthode de décomposition en quadri-Grams a été appliquée sur les lemmes afin d'obtenir une matrice de taille réduite par rapport aux mots. Pour plus de notions sur cette méthode, nous renvoyons le lecteur aux articles de [Kimbrell, 88] et de [Damashek, 95].

On produit ainsi une matrice indiquant pour chaque lemme sa fréquence dans chaque segment du texte. Nous obtenons 59 segments et 496 lemmes pour le texte A et 256 segments et 9545 lemmes pour le texte B. Nous procédons alors à la comparaison de trois méthodes de classification dont nous exposons le principe général dans ce qui suit.

a. ***La méthode des k-means*** *:* Le principe de cette méthode introduite par MacQeen en 1967 repose sur une classification par centres mobiles. Soit I un ensemble de n individus caractérisés par p mesures (variables). On suppose que l'espace R^p supportant les n individus est muni d'une distance euclidienne (ou autre) entre le i^{eme} individu et une classe l définie par:

$$D(i,l)=(\{1\leq j\leq p\}[X(i,j)-X_m(l,j)]^2)^{1/2}$$

où X(i,j) est la valeur de la $j^{ème}$ variable pour le $i^{ème}$ individu et X_m sa moyenne dans la classe l.

On détermine une partition initiale de l'ensemble des individus I en K classes par tirage pseudo-aléatoire. L'algorithme consiste ensuite à déterminer en une ou plusieurs itérations les nouveaux centres des classes d'une nouvelle partition P(n,K) induite par la réaffectation de l'ensemble I des individus dans les K classes de façon à minimiser l'erreur :

$$E[P(n,K)]=\sum_{i=1,n} D[i,l(i)]^2$$

où l(i) est la classe du i^{eme} individu, P[n,K] la partition induite par la réaffectation des individus dans les K classes et D[i,l(i)] la distance euclidienne entre l'individu i et le centre de sa classe l(i).

b. ***Le réseau connexionniste Art*** ***:*** L'idée de base du modèle ART est celle d'un système d'interaction entre 2 niveaux qui entrent en résonance mutuelle. Le système reçoit en un premier niveau N1 des stimuli qui sont envoyés mais aussi modifiés (selon une distribution et un poids particulier) au deuxième niveau N2 qui est un niveau d'archivage. Arrive donc au deuxième niveau un pattern différent de celui qui était à l'intrant. La correspondance entre le pattern prototypal et le pattern intrant est alors mesurée à l'aide d'un processus dit de résonance. Ce processus consiste à comparer les patterns intrants au pattern prototype. S'il y a correspondance, l'intrant sera alors classé avec le prototype, sinon il sera considéré comme un prototype en émergence et, il servira également comme nouveau gabarit aux autres intrants que le système introduira. Au fur et à mesure que l'apprentissage se poursuit, il y a consolidation de cette résonance [Carpenter et Grossberg 1988].

c. ***Le modèle des champs de Markov :*** Le modèle des champs de Markov, comme tout modèle probabiliste, considère les objets à classer comme des réalisations d'une famille de variables aléatoires. A nos segments, définis sur notre corpus, est associée une famille de variables aléatoires munies d'une distance (euclidienne ou autre) morpho-syntaxique et jouissant d'une propriété locale suivant un système de voisinage. Ainsi, le champs de variables aléatoires est identifié à un champs de Markov [Bouchaffra et Meunier, 95, 96]. Le détail de l'expérimentation de cette méthode est exposé dans l'article [Remaki et Meunier, JADT 98].

4. Présentation des résultats

Notons d'abord que les segments dans le texte A sont répartis de la façon suivante : Spirale (de 1 à 41 et de 46 à 60), Hydro-Québec (de 43 à 45) et le segment 42 chevauche entre les deux textes. Dans le cas du texte B, ce sont les premier et dernier segments d'Hydro-Québec qui chevauchent avec le texte de spirale.

L'analyse des résultats obtenus peut être menée selon deux points de vue : la comparaison du comportement des classifieurs quant à la classification des segments d'Hydro-Québec supposés contenir la même information sémantique, ainsi que de leur degré d'efficacité face à de gros corpus. Les trois méthodes citées ci-dessus génèrent des classes de segments qui présentent une certaine similarité lexicale entre eux. On s'attend à priori, à ce que les segments d'Hydro-Québec se retrouvent dans une même classe. Les résultats obtenus sur le texte A et pour les trois classifieurs se présentent comme suit :

Modèle de Markov **:** classe1 [1, 2, 3, 4, 5, 15, 19, 27, 31, 32] , classe2 [6, 7, 9, 17, 24], classe3 [8, 14, 38], classe4 [10, 35, 50], classe5 [11, 12, 13, 18, 23], classe6 [16, 33], classe7 [20, 22, 26, 29, 51], classe8 [21, 30], classe9 [25, 28, 34], classe10 [36, 37, 39, 48], classe11 [40, 41, **42**, 54], classe12 [**43, 44, 45**], classe13 [47, 55, 58], classe14 [49, 52], classe15 [53], classe16 [56, 57], classe17 [59]

La méthode des K-means : classe1 [5, 6, 7, 8, 9, 11, 15, 24, 26, 27, 28, 32], classe2 [3, 4], classe3 [2, 10], classe4 [**43, 44, 45**], classe5 [35, 36, 37, 39, 40], classe6 [52, 53, 55], classe7 [1, 14], classe8 [12,13], classe9 [29, 30, 31], classe10 [23, 56, 57], classe11 [17, 49, 51], classe12 [16, 18, 19, 20, 21, 22], classe13 [33, 34], classe14 [41, 48, 54], classe15 [38, **42**], classe16 [46, 47, 50, 58], classe17 [25, 59]

Le réseau connexionniste ART : classe1 [3, 32], classe2 [24, 29, 33, 38, **45,** 58], classe3 [59], classe4 [49, 51, 54], classe5 [39, 55, 56, 57], classe6 [16, 17], classe7 [34], classe8 [52], classe9 [40, 47], classe 10 [12, 37, 46, 53], classe11 [1, 35, 36, 41, **42, 43, 44**, 48], classe12 [10, 15, 21, 30, 50],
classe13 [2, 4, 5, 6, 7, 8, 9, 11, 14, 18, 19, 20, 22, 23, 25, 26, 27, 28, 31], classe14 [13], classe15 [39, 52]

En analysant ces résultats, nous constatons que les méthodes de Markov et des K-means partitionnent les segments de l'Hydro-Québec d'une façon quasi-similaire. En effet, la partition obtenue pour chacune des 2 méthodes contient une classe constituée exclusivement par les segments 43, 44 et 45. Dans les 2 cas, le segment 42 se retrouve avec des segments de Spirale. Nous pouvons expliquer ceci par le fait que ce segment contient 2 informations sémantiques différentes. Dans le cas du modèle ART, la discrimination entre les segments de l'Hydro-Québec et de Spirale n'est pas très forte. Sur le texte B, un résultat préliminaire qui nous semble intéressant est que le modèle de Markov a pu partitionner les segments du texte d'Hydro-Québec dans des classes à part. La méthode des K-means a donné également des classes propres aux segments d'Hydro-Québec, mais en a classé quelques uns avec des segments de Spirale. Ceci pourrait être du à plusieurs facteurs. La méthode des K-means dépend fortement de la classification initiale et de l'ordre d'entrée des individus à classer contrairement au modèle de Markov. La probabilité de retrouver le même n-gram dans différents segments est élevée ce qui pourrait à priori expliquer cette classification. Ceci pourrait être confirmé en comparant de tels résultats avec une classification obtenue en prenant comme lemmes les mots du texte.

5. Conclusion

Cette étude a consisté à explorer différents classifieurs sur un même texte traitant de sujets de nature sémantique différente. Ceci nous a permis de procéder à une comparaison des classifications

obtenues. Cependant, en sus de ces résultats une analyse plus approfondie au niveau des mots dans une même classe, pourrait expliquer plus profondément les différences dans les résultats. Nous envisageons également de compléter cette étude en comparant les classifications obtenues, d'une part sur les N-grams et d'autre part sur les mots. Ceci nous permettra de *"voir"* dans quel cas nous accédons mieux au contenu sémantique. En effet, une bonne classification s'avère très importante du point de vue terminologie, extraction de connaissances ou plus particulièrement navigation hypertextuelle.

Bibliographie

Balpe, J. P., Lelu, A., Papy, F., & I, S. (1996). *Techniques avancées pour l'hypertexte.* Paris : Hermes.

Burr, D. J. (1987). "Experiments with a connectionnist text reader". IEEE First International Conference on Neural Networks, San Diego, 717-24

Carpenter, G. & Grossberg, G. (1991). "An Adaptive resonnance Algorithm for Rapid Category Learning and Recognition". Neural Networks 4, 493-504.

Church, K., Gale, W., Hanks, P., & Hindle, D. (1989). "Word Associations and Typical Predicate-Argument Relations". *International Workhop on Parsing technologies*,Carnegie Mellon University, Aug. 28-31,

Church, K. W., & Hanks, P. (1990)."Word association norms, mutual informaton, and lexicography". *Computational Linguistics 16* , 22-29.

Delisle, S. (1994). Text Processing without a priori domain knowledge: semi automatic linguistic analysis for incremental knowledge acquisition. PH Thesis, Ottawa University. :

Garnham, A. (1981). "Mental models and representation of texts". Memory and Cognition 9 (560-565),

Grefenstette, G. (1992). "Sextant: Exploring Unexplored Contexts for Semantic Extraction from Syntactic Analysis". Proc of the 30th Annual Meeting fo the ACL 324- 326,

Grefenstette, G. (1992). "Use of syntactic Context to Produce Term Association Lists for Text Retrieval". Proc of SIGIR 92 ACM, Copenhagen, june 21-24,

Grossberg, S. , & Carpenter, S. (1987). "Self Organization of Stable Category Recognition Codes for Analog Input Patterns". Applied Optics 26, 4919- 4930.

Jacobs, P. , & Zernik. U. (1988). "Acquiring Lexical Knowledge from Text A case Study". Proceedings of AAA1 88 (St Paul. Min.),

Kahonen, T. (1982). "Clustering, taxonomy and topological Maps of Patterns". IEEE Sixth International Conference on Pattern Recognition, 114-122

Lebart, L. , & Salem, A. (1988). Analyse statistique des données textuelles. Paris: Dunod.

Lelu, A. (1995). "Hypertextes: la voie de l'analyse des données". In L. Bolasco..S L ,A.Salem (Ed.), *Anilisi statistica dei dati testuali vol2. (pp.* 85-96). Rome: CISU.

Lin, X. , Soergel, D. , & Marchionini, G. (1991). "A Self Organizing Semantic Map for Information Retrieval". SIGIR 91, Chicago, Illinois,

Meunier,J.G (1996) Théorie cognitive:son impact sur le traitement de l'information textuelle.in V.Riale et D. *Fisette Penser L'esprit ,Des sciences de la cognition a une philosophie cognitive.* Presses de Université de Grenoble. 1996 289-305

Moulin B, & Rousseau, D. (1990). "Un outil pour l'acquisition des connaissances a partir de textes prescriptifs". ICO, Québec 3 (2), 108-120.

Recoczei, S. , & E. P. O, P. (1988). "Creating the Domain of Discourse: Ontology and Inventory". In J. &. B. G. Boose (Ed.), Knowledge Acquisition Tools for Experts and Novices. Academic Press:

Regoczei, S. , & Hirst, G. (1989). On extracting knowledge from Text. Modeling the Architecture of Language Users. (TR CSRI 225). Computer Systems Research Institute University of Toronto.

Salton, G. (1988). "On the Use of Spreading Activation". Communications of the ACM vol 31 (2),

Salton, G. , Allan, J. , & Buckley, C. (1994). "Automatic Stucturing and Retrieval of Large Text File". Communications of the ACM 37 (2), 97-107.

Tapiero, I. (1993). Traitement cognitif du texte narratif et expositif et connexionnisme: expérimentations et simulations. in Université de Paris VIII,

Thrane, T. (1992). "Dynamic Text Comprehension". In J. O. S. Jansen H Prebensen, T. Thrane (Ed.), Copenhaguen: Museum Tuscalanum Press.

Veronis, J. , Ide, N. M. , & Harie, S. (1990). "Utilisation de grands réseaux de neurones comme modèles de représentations sémantiques". Neuronimes,

Virbel, J. (1987)."L'apport de conniasances linguistiques à l'interprétation des structures textuelles". *Structure des documents, Bigre++Globule 53* , 77-97.

Virbel, J. E., F. Pascual, E. (1992). *La lecture assisfée par ordinateur,Raport de recherche.* Toulouse: Laboratoire IRIT.

Virbel, J. (1993). "Reading and Managing Texts on the Bibliothèque de France Stations". In P. Williams, M. (1990). " Connectionist Models and Information Retrieval". 25, 209-259.

Young, T. , & Calvert, T. (1987). Classification, Estimation, and Pattern Recognition. Amsterdam: Elsivier.

Zarri, G. P. (1990). "Représentation des connaissances pour effectuer des traitements inférentiels complexes sur des documents en langage naturel.". In Office de la langue française (Ed. Les industries de la langue. Perspectives 1990. Gouvernement du Québec.