Le modèle "monomaniaque" un modèle statistique simple pour l'analyse exploratoire d'un corpus de textes

Fabrice Clérot, Olivier Collin, Olivier Cappé, Eric Moulines

Le modèle "monomaniaque"
un modèle statistique simple pour l'analyse exploratoire d'un corpus de textes

F. Clérot[1], O. Collin[1], O. Cappé[2], E. Moulines[2]
[1]*France Télécom R&D, 2 avenue Pierre Marzin, 22300 Lannion, France*
[2]*GET/ENST et CNRS/LTCI, 46 rue Barrault, 75013 Paris, France*
{fabrice.clerot,olivier.collin}@francetelecom.fr, {cappe,moulines}@enst.fr

RÉSUMÉ. *Regrouper les éléments d'un corpus de textes en segments thématiquement apparentés est un problème d'analyse exploratoire complexe. On explore dans cette communication les performances d'un modèle statistique simple, le modèle "monomaniaque".*

On décrit le modèle et sa procédure d'ajustement puis on illustre sa performance sur un problème de segmentation d'un corpus de textes anglais relatifs au CKM ("Customer Knowledge Management")..

ABSTRACT. *The automatic clustering of text segments into thematically homogeneous groups is a difficult problem. In this paper, we study the performance of a simple probabilistic model, the "monomaniac" model.*

We first describe the model and the related EM-based estimation procedures; an application of this model to a English corpus of texts imported from the CKM ("Customer Knowledge Management") literature is then presented.

MOTS-CLÉS : *analyse de données textuelles, clustering, modèle de mélange.*

KEYWORDS: *textual data analysis, clustering, mixture models.*

Actes de CIFT 2004, pages 45 à 57

1. Introduction

Le clustering est une des techniques de base de l'analyse de données exploratoires. Les groupes homogènes obtenus apportent une compréhension plus synthétique des données, en permettent une visualisation plus naturelle et peuvent servir de base à la construction d'autres démarches (classification supervisée par exemple). Moins classique en analyse de données textuelles, ce problème a récemment reçu une grande attention et des techniques ont été développées qui ambitionnent d'apporter dans le domaine de l'analyse de données textuelles les mêmes avantages.

Dans ce domaine, l'ambition du clustering revient à construire ex-nihilo un plan de classement à partir d'un corpus de texte supposé représentatif d'un domaine. Ce problème se rattache à l'analyse exploratoire car on suppose qu'il n'existe pas a priori de plan de classement prédéfini, ni sous forme de description sémantique des rubriques, ni sous forme d'exemples étiquetés au préalable.

Outre des problèmes techniques liés aux particularités des données textuelles, le clustering de documents se heurte au problème de la définition de l' homogénéité des clusters : en analyse de données "traditionnelle", cette homogénéité est jugée au sens d'une métrique; pour des documents, il est clair que "homogénéité" signifie "homogénéité sémantique", une notion plus difficile à quantifier. En dernier ressort, c'est à l'analyste de décider ce qui lui paraît ou non pertinent dans les résultats des analyses et d'agir sur les paramètres à sa disposition pour corriger certains comportements. Plus encore qu'en analyse de données traditionnelle, le clustering en analyse de données textuelles doit être considéré comme un outil d'exploration du corpus à la disposition de l'analyste et non comme une technique "presse-bouton" fournissant un résultat indiscutable.

Parmi les apports attendus de ce type d'analyse, on peut citer l'aide à la génération automatique d'ontologies pour la gestion de la connaissance et de l'information ("knowledge management") ou le tri rapide de corpus de textes non étiquetés, pour la veille technologique.

Ci-dessous, Section 2, on décrit le modèle. La Section 3 précise notre démarche d'analyse exploratoire par rapport à l'état de l'art actuel. La méthode d'ajustement des paramètres est décrite en Section 4. On donne enfin Section 5 un exemple d'utilisation de ce modèle pour l'analyse exploratoire d'un corpus lié à un domaine assez étroit, la gestion de la relation client.

2. Présentation du modèle monomaniaque

L'application des méthodes statistiques aux données textuelles suppose une représentation numériques des données: un document est réduit à un vecteur de nombres qui représentent le nombre d'occurrences (normalisé ou non) de "mots" d'un dictionnaire défini à l'avance dans ce document. Le corpus de documents est donc représenté comme un ensemble de points dans un espace vectoriel de grande dimension (autant que de "mots" dans le dictionnaire). C'est la densité de probabilité sous-jacente à cet ensemble d'observations qu'on cherche à estimer.

On introduit les notations suivantes : les "mots" sont indexés par $w = 1,\Lambda\ ,W$, les documents sont indexés par $d = 1,\Lambda\ ,D$ et le corpus des documents est donc représenté par une matrice de comptes $C = \left(c_d^w\right)_{d=1\Lambda\ D}^{w=1\Lambda\ W}$.

Le plan de classement visé par le modèle monomaniaque est le plus simple possible, constitué de T "thèmes" mutuellement exclusifs; dans la suite, les thèmes sont indexés par $t = 1,\Lambda\ ,T$.

Les hypothèses qui sous-tendent le modèle sont les suivantes:

1. un document d est associé a priori à un thème t et un seul (d'où le surnom de "monomaniaque")
2. conditionnellement à un thème t , les comptes des mots sont distribués indépendamment
3. la distribution des thèmes suit une loi multinomiale $Mult\left(1; p_1,\Lambda\ , p_T\right)$
4. les comptes du mot w conditionnellement au thème t suivent une loi de Poisson de paramètre $\lambda^w(t)$

Les paramètres qui définissent le modèle sont donc:

- T le nombre de thèmes
- $\left(p_t\right)_{t=1\Lambda\ T}$ les probabilités des thèmes (T paramètres)
- $\left[\lambda^w(t)\right]_{t=1\Lambda\ T}^{w=1\Lambda\ W}$ les paramètres des lois de Poisson régissant les comptes de mots conditionnellement aux thèmes ($W \times T$ paramètres)

Les deux premières hypothèses sont constitutives du modèle monomaniaque et permettent d'exprimer simplement la probabilité d'un document d sous la forme d'un mélange de lois :

$$\Pr\left[\left(c_d^w\right)^{w=1\Lambda\ W}\right] = \sum_{t=1}^{T} \Pr(t)_t \prod_{w=1}^{W} \Pr\left(c_d^w | t\right) \qquad (1)$$

Les deux dernières hypothèses spécifient seulement la forme des lois du mélange; le choix de lois de Poisson pour représenter la distribution des comptes répond à un souhait de simplicité technique maximale et est cohérent avec l'observation très générale de l'extrême petitesse des comptes observés: la matrice représentative du corpus est une matrice très "creuse".

3. La démarche d'analyse exploratoire

Ce type de modèle a été introduit dans (Nigam et al, 1999) ou (Nigam et al, 2000) (avec un choix différent pour la loi des comptes; les auteurs utilisant une loi multinomiale). Les résultats présentés dans (Blei et al, 2003) montrent que l'hypothèse (1) ci-dessus est trop restrictive pour pouvoir être appliquée directement à un corpus de documents: on comprend que la thématique

puisse changer au fil d'un document, surtout si ce document est long. Les auteurs proposent un modèle plus complexe permettant à un document d'appartenir à plusieurs thèmes (Latent Dirichlet Allocation, LDA). Ce modèle apparaît comme une généralisation du modèle pLSI (probabilistic Latent Semantic Indexing) (Girolami et Kaban, 2003) introduit dans (Hoffman, 1999). Dans les deux cas, les auteurs cherchent à capturer en une seule étape la structure statistique du corpus, quitte à utiliser pour cela des modèles difficiles à interpréter (pLSI) ou à ajuster (LDA).

Notre approche est différente en ce sens que le modèle ci-dessus ne constitue que la première partie d'une démarche d'analyse exploratoire: *le modèle monomaniaque est appliqué non pas au corpus des documents eux-mêmes mais au corpus des paragraphes contenus dans ces documents*. Autant l'hypothèse "mono-thématique" paraît réductrice au niveau d'un document tout entier, autant cette hypothèse nous paraît naturelle au niveau d'un paragraphe.

Cette première étape franchie, on pourra ensuite associer une représentation thématique à chaque document en associant à chaque document les thèmes de ses paragraphes. Une représentation simple pourra être le vecteur à T composantes du nombre d'apparitions de chaque thème dans le document; d'autres représentations plus complexes sont possibles comme la succession des thèmes dans le document.

A partir de cette représentation, on pourra reprendre une nouvelle phase d'analyse exploratoire dans un espace de dimension réduite où on pourra appliquer l'arsenal des techniques plus traditionnelles d'analyse de données. Cette technique d'analyse exploratoire comportant plusieurs étapes d'agrégation a déjà été exploitée dans des domaines très différents (voir (Clérot et Fessant,2004), par exemple).

Pour ce type de démarche, il est important de construire à chaque étape des groupes aussi clairement tranchés que possible et c'est en cela que l'hypothèse "mono-thématique" est essentielle: en imposant cette contrainte comme un a priori du modèle, on espère pouvoir a posteriori attribuer la majorité des documents à un thème seulement sans ambiguïté.

4- Estimation des paramètres du modèle

On estime le modèle de mélange (1) en maximisant la vraisemblance du modèle par la méthode EM (Titterington, 1985). Pour cela, on introduit comme a priori sur la distribution des paramètres à estimer:

- la loi de Dirichlet de paramètres $(\theta,\Lambda\ ,\theta)$, conjuguée de la loi multinomiale;
- la loi Gamma de paramètres (α,β), conjuguée de la loi de Poisson

On emploie la même valeur des paramètres pour toutes les lois.

L'emploi des lois conjuguées permet de donner une forme analytique simple aux formules de réestimation (Denison et al, 2002). On obtient :

$$p_t = \frac{(\theta - 1) + \sum_{d=1}^{D} \gamma_d(t)}{T(\theta - 1) + \sum_{t=1}^{T} \sum_{d=1}^{D} \gamma_d(t)} \tag{2}$$

$$\lambda^w(t) = \frac{(\alpha - 1) + \sum_{d=1}^{D} c_d^w \gamma_d(t)}{\beta + \sum_{d=1}^{D} \gamma_d(t)} \tag{3}$$

On a introduit:

$$\gamma_d(t) = \frac{p(t) \times \exp(-\sum_{w=1}^{W} \lambda^w(t)) \times \prod_{w=1}^{W} \lambda^w(t)^{c_d^w}}{\sum_{s=1}^{T} p(s) \times \exp(-\sum_{w=1}^{W} \lambda^w(s)) \times \prod_{w=1}^{W} \lambda^w(s)^{c_d^w}} \tag{4}$$

dont l'interprétation est la probabilité a posteriori pour le document d d'appartenir au thème t.

On initialise l'algorithme en partant d'un tableau $\left[\gamma_d(t)\right]_{d=1\Lambda D}^{t=1\Lambda T}$ tiré aléatoirement entre 0 et 1 puis convenablement normalisé.

Les paramètres des lois conjuguées peuvent permettent d'influencer le comportement de la solution; en particulier, choisir le paramètre θ inférieur à 1 revient à préférer une distribution très "tranchée" des probabilités des thèmes avec des thèmes ayant une probabilité significative et les autres des probabilités quasi-nulles. On a choisi de travailler avec $\theta = 1/2$, ("a priori de Jeffreys"), sans voir de différence notable avec le choix $\theta = 1$ (a priori uniforme).

Pour les paramètres des lois Gamma, on a fixé la moyenne à la moyenne des comptes observés sur l'ensemble du corpus et on a choisi une variance très grande par rapport à la variance observée sur l'ensemble du corpus.

A la fin du processus d'estimation, on obtient les valeurs des paramètres du modèle et on peut remarquer que connaissant ces paramètres, (4) permet d'estimer la probabilité qu'un document quelconque appartienne à un thème pourvu que ce document soit représenté comme un vecteur de comptes $\left(c_d^w\right)_{w=1\Lambda W}$ sur la base des mots du vocabulaire. *L'expression (4) permet donc de classifier aussi les nouveaux documents*, pas seulement les documents du corpus.

On remarque encore que le nombre de thèmes T est fixé a priori par l'analyste; il est possible de sélectionner la meilleure valeur en analysant la variation de la vraisemblance du modèle en

fonction de T et en choisissant une valeur dans la région où ajouter des thèmes n'augmente plus significativement la vraisemblance du modèle.

5- Illustration des performances du modèle

5.1- Le corpus étudié et les prétraitements

Ce corpus regroupe des documents en langue anglaise relatifs à la gestion de la relation client, CKM (Customer Knowledge Management). La construction automatique à partir des ressources traditionnelles d'analyse des langues naturelles d'un plan de classement pour un corpus de documents représentatif d'un domaine aussi étroit est un problème difficile car ces ressources plutôt "généralistes" n'ont souvent pas un grain d'analyse assez fin pour résoudre des catégories à l'intérieur du domaine. L'acquisition manuelle préalable de ces ressources "spécialisées" est coûteuse en temps et nécessite la collaboration sur toute cette période d'un spécialiste du traitement des langues naturelles et d'un spécialiste du domaine. Dans une activité comme la veille technologique qui, par principe, multiplie les recherches dans des domaines d'activité très pointus, cette acquisition manuelle ne peut pas être envisagée durablement.

Le corpus est constitué d'un ensemble de documents HTML, de tailles très variables, « nettoyés » de leur balises, découpés en mots, phrases et paragraphes par un traitement classique de segmentation. Un étiquetage basée sur une analyse syntaxique projette ensuite les segments sur un ensemble de lemmes associés à des catégories syntaxiques fines. Des essais préliminaires ont montré qu'un regroupement des lemmes suivant ces sous-catégories n'est pas utile pour notre problème. Nous avons donc conservé un niveau de représentation plus grossier correspondant aux catégories usuelles (nom, verbe, adjectif…). Les lemmes des noms propres sont les noms propres eux-mêmes. Comme on l'a indiqué plus haut, le modèle monomaniaque est appliqué au corpus des paragraphes. Des identificateurs uniques sont associés à chaque lemme, document et paragraphe, afin de constituer les matrices de comptages paragraphes-lemmes.

Les mots grammaticaux de l'anglais ainsi que les mots très fréquents sont supprimés. Parmi les lemmes restants, on a retenu seulement les 1000 plus fréquents pour la description du corpus. De même, on a ignoré dans le corpus les paragraphes comportant moins de 10 lemmes. Le corpus comporte finalement 12489 paragraphes différents.

Une première passe de l'algorithme a révélé des groupes répondant visiblement à des idiosyncrasies du corpus; le corpus contenant de nombreux extraits de rapports du Gartner Group ou de Datamonitor, on a trouvé des groupes essentiellement constitués de titres de ces documents, caractérisé par la présence constante des lemmes "gartner" ou "datamonitor". Ce comportement est à la fois satisfaisant sur le plan des principes (le groupe est effectivement homogène et bien caractérisé) et peu satisfaisant sur le plan applicatif. On a supprimé manuellement certains lemmes qui donnaient lieu à ce type de regroupement.
Les résultats présentés ci-dessous concernent un corpus de 12489 paragraphes représentés sur 973 lemmes.

5.2 Visualisation et interprétation des résultats

On a choisi de présenter les résultats obtenus en fixant le nombre de thèmes à $T = 20$. On obtient 20 thèmes qui ont tous des probabilités significatives (Figure 1).

La Figure 2 montre la distribution cumulée inverse pour tous les paragraphes de $\max_{t=1..20}[\gamma_d(t)]$, le poids du thème le plus probable pour chaque paragraphe. On y voit en particulier qu'a posteriori, pour 75% des paragraphes, ce maximum est supérieur à 0.5 (les 19 autres thèmes se partageant les 0.5 restants) et que pour la moitié des paragraphes, ce maximum est supérieur à 0.9 (les 19 autres thèmes se partageant les 0.1 restants). On peut conclure que l'a priori "mono-thématique" est bien suivi par la distribution a posteriori et qu'on a obtenu un plan de classement permettant l'attribution sans ambiguïté d'une grande majorité des paragraphes à un thème et un seul.

A la différence de l'analyse supervisée où les catégories sont données à l'avance, la pertinence d'une analyse exploratoire repose sur sa capacité à permettre une meilleure appréhension des données. Dans le cas du clustering, il faut donc encore pouvoir interpréter les groupes qui ont été formés par l'analyse. Ce travail d'interprétation stricto sensu n'est pas du ressort de l'analyste de données mais d'un expert du domaine (ici le CKM) en collaboration avec l'analyste. Un des objectifs de l'analyste de données est de présenter à l'expert du domaine une vision aussi synthétique que possible des résultats afin de lui permettre une interprétation rapide.

Le fait que chaque groupe soit représenté par un grand nombre (973) d'intensités de Poisson est évidemment insuffisant pour l'interprétation par l'expert du domaine; on introduit ci-dessous une visualisation qui permet d'interpréter assez naturellement les résultats. Pour chaque thème:

- on relève les "mots-clés", les mots dont la fréquence dans le thème relativement à leur fréquence dans tout le corpus est la plus grande;
- on calcule la fréquence de co-occurrence de ces mots-clés dans le thème relativement à leur fréquence dans tout le corpus; cette fréquence est interprétée comme une similarité entre mots-clés et cette information de similarité est visualisée sous forme d'une projection 2D par multi-dimensional scaling (cette projection non linéaire vise à respecter en basse dimension les relations de voisinage existant dans l'espace de départ (Cox, 2001)).

Un exemple d'une telle projection est donné ci-dessous (Figure 3; on a conservé les 30 premiers mots-clés). La taille des points indique le rang des mots-clés, la relation de similarité est codée en niveau de gris sur les arêtes; la couleur plus ou moins forte du point code la somme des similarités associées au mot-clé correspondant. Cette projection concerne le thème le plus répandu dans le corpus; on y trouve en effet tous les mots-clés qu'on pouvait attendre dans un corpus relatif à la gestion de la relation client. On peut remarquer que le mot-clé le plus central dans cette projection ("customer") peut être interprété comme celui ayant le plus de similarité avec tous les autres; ce n'est pas forcément le mot clé le plus saillant ("segmentation"). Les paragraphes associés à ce thème sont des paragraphes qui traitent de façon générale de la gestion de la relation-client et de ses objectifs.

On donne un autre exemple Figure 4 ; l'interprétation de ce thème est assez facile, il s'agit de paragraphes traitant de reconnaissance et de traitement automatique de la parole.

5.3- Récapitulation

L'interprétation de tous les thèmes n'a toutefois pas été possible; le

Tableau 1 résume les interprétations. On peut retenir que:

- tous les thèmes les plus importants ont été identifiés;
- certains thèmes de faible probabilité sont également identifiables, par conséquent, il n'y a pas lieu de ne considérer comme "bien formés" que les seuls thèmes de forte probabilité;
- les thèmes clairement identifiés représentent au total plus de 60% des paragraphes.

Conclusion

On a présenté dans cette communication un modèle statistique simple permettant le clustering non supervisé d'un corpus de documents sous l'hypothèse que chaque document puisse être attribué à un thème et un seul (modèle "monomaniaque").

A la différence d'autres études sur le sujet, cette étape de clustering s'intègre dans une démarche complète d'analyse exploratoire et on applique ce clustering au corpus des paragraphes et non à celui des documents. A ce niveau, l'hypothèse "mono-thématique" apparaît justifiée.

On n'a présenté ici que les résultats relatifs au seul clustering des paragraphes; on a pu constater que le clustering obtenu permettait d'attribuer la grande majorité des paragraphes à un thème seulement et qu'une visualisation assez simple permettait à un expert du domaine d'interpréter une majorité des thèmes.

Bibliographie

K. Nigam, J. Lafferty, A. McCallum, "Using maximum entropy for text classification", *IJCAI, Workshop on Artificial Intelligence for Information Filtering*, 61-67, 1999.

K. Nigam, A. McCallum, S. Thrun, T. Mitchell, "Text classification from labelled and unlabelled data using EM", *Machine Learning*, 39(2/3):103-134, 2000.

D. Blei, A. Ng, M. Jordan, "Latent Dirichlet allocation", *Journal of Machine Learning Research*, 3(5):993-1022, 2003.

M. Girolami, A. Kaban, "On an equivalence between pLSI and LDA", in *Proceedings of the 26th Annual ACM Conference on Research and Development in Information Retrieval*, SIGIR, 433-434, 2003.

T. Hoffman, "Probabilistic latent semantic indexing", in *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, 50-57, 1999.

F. Clérot, F. Fessant, "From IP port numbers to ADSL customer segmentation: knowledge aggregation and representation using Kohonen maps", *Data Mining IV* (Rio de Janeiro), Eds Ebecken, Brebbia, Zanasi, WIT Press (2004).

D. Titterington, A. Smith, U. Makov, *Statistical analysis of finite mixture distributions*, Wiley (1985).

D. Denison, C. Homes, B. Mallick, A. Smith, *Bayesian methods for nonlinear classification and regression*, Wiley (2002).

T. Cox, M. Cox, *Multidimensional scaling*, Chapman and Hall, 2001.

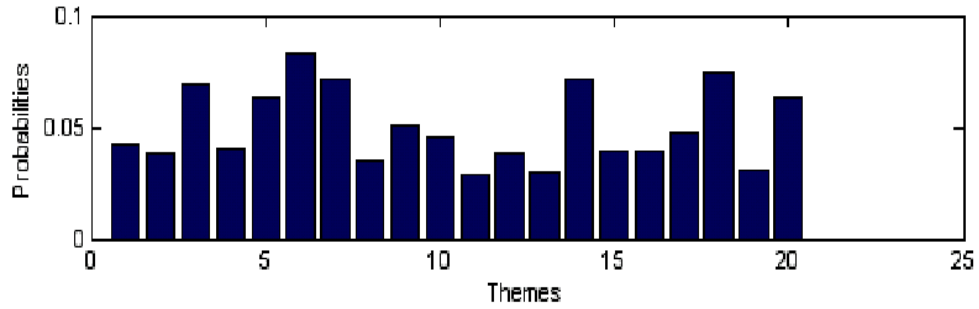

Figure 1: probabilités des thèmes

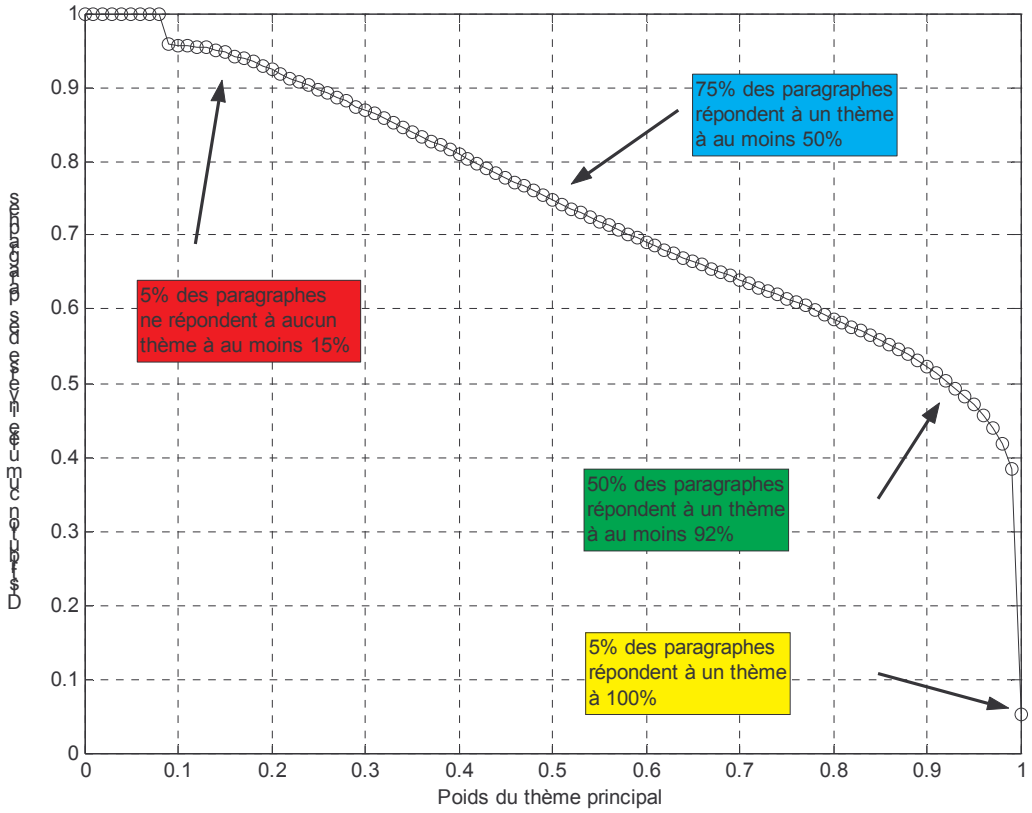

Figure 2: distribution cumulée inverse de la contibution du thème principal

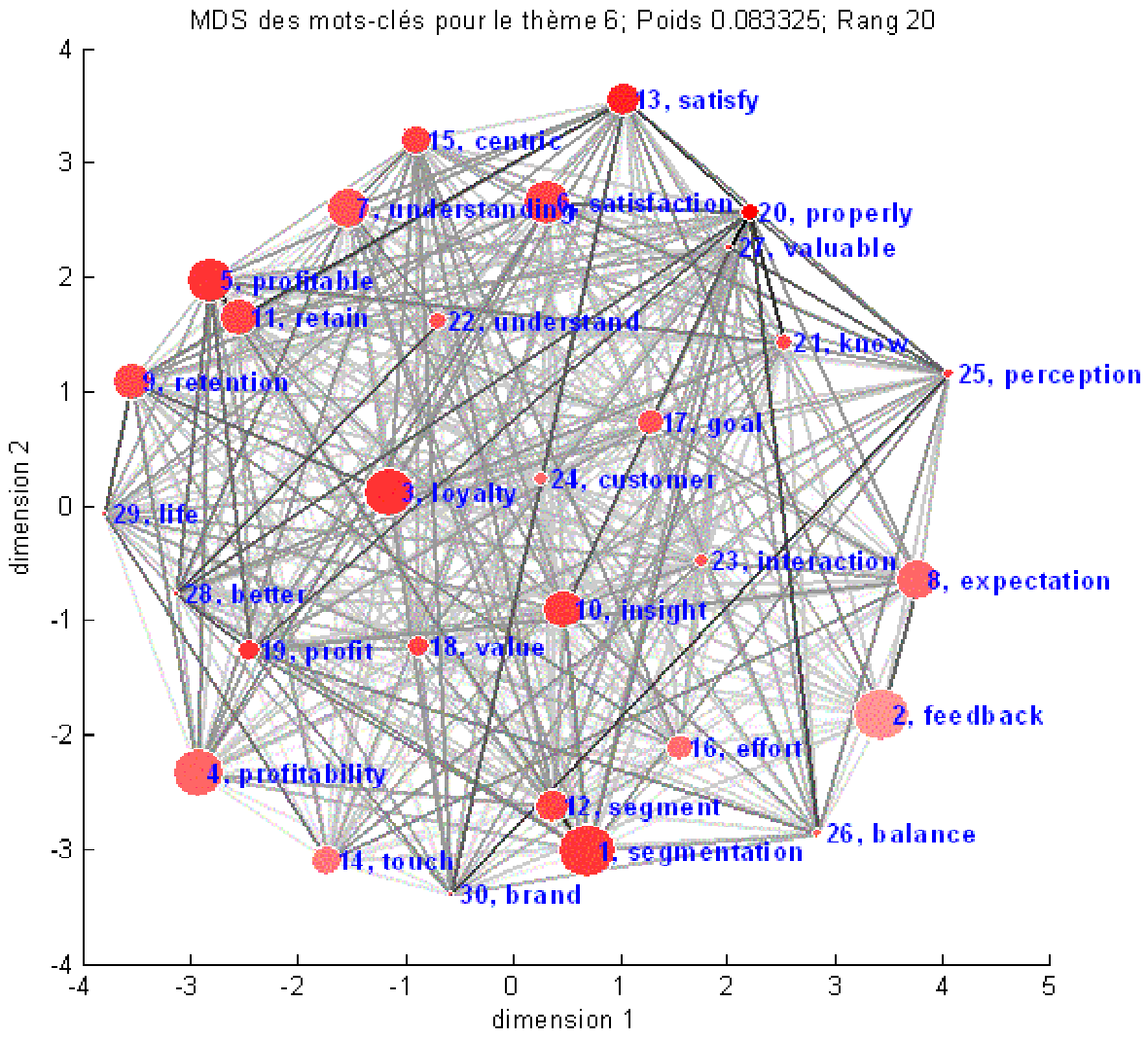

Figure 3: visualisation du thème principal

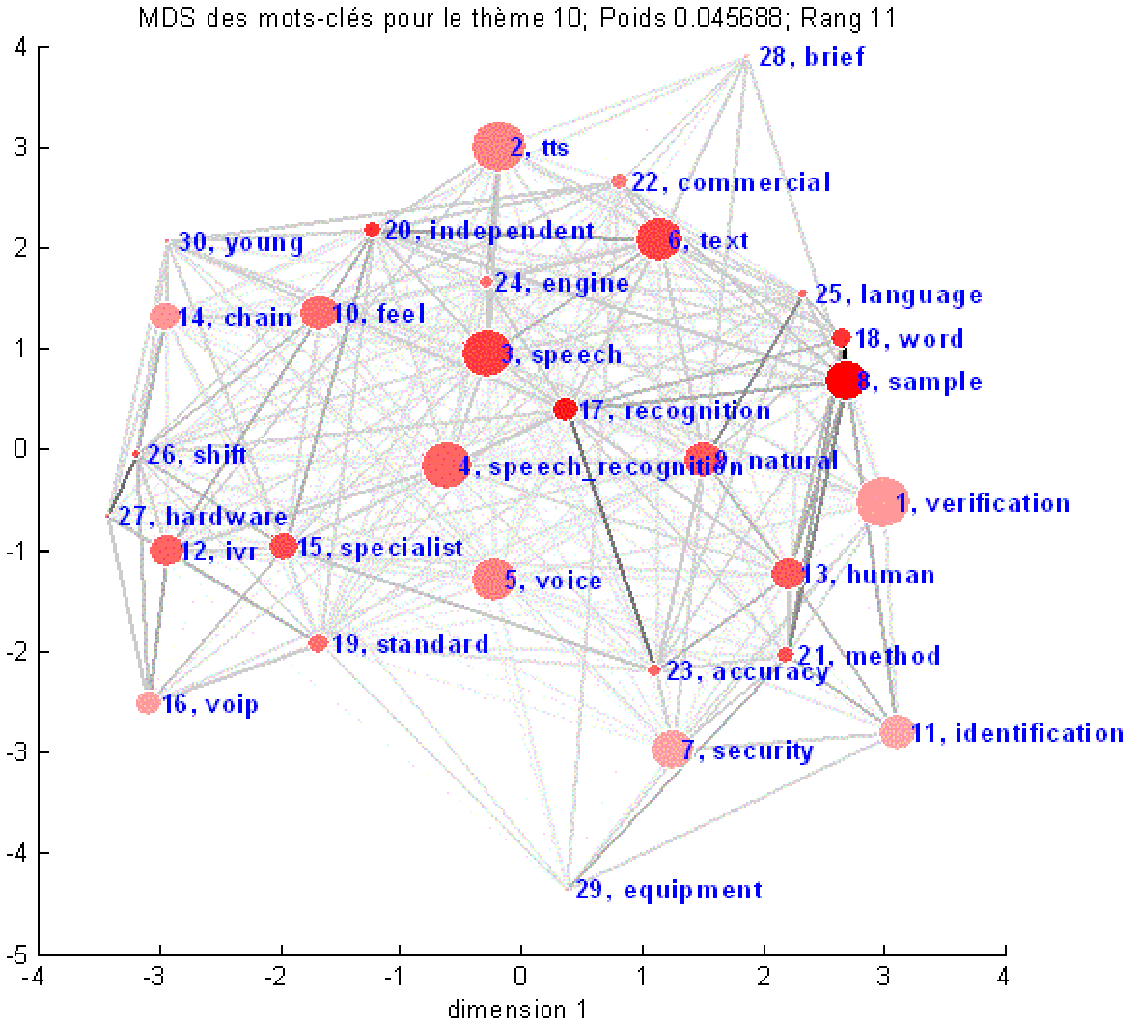

Figure 4: projection du thème "reconnaissance et traitement de la parole"

	INTERPRETATION		OUI	Douteux	NON
1	Présentation générale du CKM	**8.3%**			
2	Conduite d'un projet CKM		**7.5%**		
3	Vente en ligne	**7.2%**			
4	Personnalisation des contenus	**7.1%**			
5	Centre d'appels	**6.9%**			
6	Comparaison géographique	**6.3%**			
7	Banque, assurance		**6.3%**		
8	?				(5.0%)
9	?				(4.7%)
10	Traitement de la parole	**4.6%**			
11	?				(4.2%)
12	PRM (?)			4.0%	
13	?				(3.9%)
14	?				(3.9%)
15	?				(3.8%)
16	Mobilité	**3.8%**			
17	?				(3.5%)
18	Idiosyncrasie REUTERS			3.1%	
19	Configurator (?)		3.0%		
20	Condition d'usage des suites de CKM	**2.9%**			

Tableau 1: récapitulation des interprétations