île Rousse 2005
Journée sur les systèmes d'information élaborée

UN NOUVEL OUTIL DE CLASSIFICATION NON SUPERVISEE DE DOCUMENTS POUR LA DECOUVERTE DE CONNAISSANCES ET LA DETECTION DE SIGNAUX FAIBLES : RARES TEXT™

Julien Ah-Pine(*), Julien Lemoine(*), Hamid Benhadda(*)
julien.ah-pine, julien.lemoine, hamid.benhadda@fr.thalesgroup.com

(*) : Thales Land and Joint / CeNTAI (Centre des Nouvelles Technologies de l'Analyse de l'Information)
160, boulevard de Valmy – BP 82 – 92704 Colombes Cedex – France

Mots clés : classification non supervisée, analyse relationnelle, text-mining, classification conceptuelle, signaux faibles, découverte de connaissances

Résumé
Nous présentons les fonctionnalités de RARES Text, outil de classification non supervisée de documents développé par Thales Land & Joint. Les caractéristiques majeures de cet outil sont les suivantes :

- moteur de classification linéaire permettant de prendre en compte des corpus de grande taille,
- classification automatique sans fixation a priori d'un nombre de classes,
- processus de classification itératif permettant la représentation du résultat selon plusieurs degrés de finesse à l'aide d'une hiérarchie conceptuelle.

Cet outil est intégré à des solutions et des plate-formes Thales orientées Défense et Sécurité.

1. Introduction

L'analyse de données textuelles connaît un intérêt grandissant depuis plusieurs années. Le développement des moyens de communications a engendré une utilisation croissante du texte comme support de l'information. Les données textuelles numérisées constituent ainsi jusqu'à 80% des flux d'informations stockées quotidiennement dans les entreprises. Dès lors, un besoin d'outils capables de gérer, d'organiser et d'analyser en profondeur et en des temps raisonnables cette masse de données s'impose. Ces outils sont notamment indispensables dans des domaines tels que la gestion de clientèle, l'analyse des résultats d'enquêtes, la veille stratégique, sanitaire et scientifique, etc…

Dans le cadre de cet article, nous nous intéressons à la méthodologie de l'analyse relationnelle qui permet de classifier de manière non supervisée des données textuelles. L'objectif est de constituer au sein d'un corpus, des groupes de documents dont les profils sont les plus similaires entre eux et les plus dissimilaires avec les profils des documents des autres groupes.
Cette méthodologie, à l'inverse des méthodes de classification supervisée (ou catégorisation) ne présuppose pas au départ une connaissance a priori de la structure du corpus.
Comme les autres méthodes de classification non supervisée, elle se base sur la définition d'une mesure de similarité (ou de dissimilarité) entre deux documents et d'un critère à optimiser, en l'occurrence le critère de Condorcet.
La mesure de similarité entre deux documents permet de quantifier le degré de proximité de leurs « thématiques » sous-jacentes. Nous supposons ici, que l'information contenue dans un corpus est représentée dans une matrice rectangulaire T binaire de présence/absence[1] de terme général T_{ij} ou chaque ligne correspond à un document et chaque colonne à une unité lexicale. Si dans un corpus, il y a N documents et M unités lexicales, nous avons donc $\forall i = 1,...,N$, $\forall j = 1,...,M$:

$$T_{ij} = \begin{cases} 1 & \text{si présence de l'unité lexicale } j \text{ dans le document } i \\ 0 & \text{si absence de l'unité lexicale } j \text{ dans le document } i \end{cases}$$

L'application qui consiste à transformer un document en un vecteur binaire s'appuie sur un étiquetage morpho-synthaxique dont nous ne parlerons pas ici.

Dans cet article, nous rappelons les méthodes statistiques classiques de classification non supervisée pour insister plus particulièrement sur les avantages qu'apporte l'analyse relationnelle. Nous présentons succinctement les caractéristiques de l'heuristique basée sur cette méthode intégrée dans RARES Text. Nous expliquons en quoi cet outil permet une aide à la découverte de connaissances et à la détection des signaux faibles.

2. Rappels

2.1. La nature combinatoire du problème de la classification non supervisée d'un ensemble d'objets

La classification non supervisée d'un ensemble d'objets est un problème hautement combinatoire. En effet, le nombre de partitions possibles $P_{n,k}$ de n objets en k classes est donné par le nombre de Stirling de deuxième espèce :

$$P_{n,k} = \frac{1}{k!}\sum_{i=1}^{k} C_k^i (-1)^{k-i} i^n$$

[1] Il est possible de prendre en compte des fréquences d'unités lexicales dans les documents. Toutefois, nous nous ramenons au cas binaire (au sens d'un tableau disjonctif) grâce à des techniques de discrétisation.

A titre illustratif, si nous devions trouver une partition optimale de 15 objets en 7 classes, il faudrait énumérer et évaluer la qualité (au sens d'un critère particulier) de un peu moins de 5 millions de partitions possibles !

$$P_{15,7} = 4\,729\,725$$

Si l'on devait envisager toutes les partitions possibles sans fixation de nombre de classes[2], il faudrait énumérer un nombre de cas égal au nombre de Bell :

$$B_n = \sum_{k=1}^{n} P_{n,k} = e^{-1} \sum_{k=1}^{\infty} \frac{k^n}{k!}$$

Ce nombre revient à envisager toutes les partitions possibles de la plus fine où tous les objets sont isolés, à la plus dense, où tous les objets sont regroupés dans une classe unique.
A titre d'exemple, le nombre de partitions possibles dans une population de 15 objets est de l'ordre de 1,4 milliards[3] !

$$B_{15} = 1\,382\,958\,545$$

Nous voyons que le nombre de partitions possibles sans fixation du nombre de classes est largement supérieur au nombre de partitions possibles avec fixation du nombre de classes. Cependant, fixer un nombre de classes ne simplifie pas le problème de la classification non supervisée pour autant. En effet, le problème reste malgré tout hautement combinatoire.

S'il fallait énumérer et évaluer la qualité de tous les cas possibles afin de déterminer la meilleure partition d'un ensemble de n objets, nous voyons que le problème de la classification non supervisé est un problème hautement coûteux en temps de calculs. Il existe plusieurs méthodologies statistiques permettant de déterminer une ou plusieurs bonnes partitions sans passer par l'énumération de toutes les partitions possibles. Certaines de ces méthodologies proposent de traiter le problème de classification non supervisée avec fixation du nombre de classes, d'autres sans fixation du nombre de classes, d'autres encore proposent une hiérarchie de partitions à nombre de classes variables. Nous rappelons succinctement ces méthodologies dans ce qui suit.

2.2. Rappels des méthodes classiques en classification non supervisée

Les méthodes classiques de classification non supervisée se décline en deux familles : la famille des méthodes de classification hiérarchique et celle des méthodes de classification non hiérarchique.

2.2.1. Les méthodes hiérarchiques

Ces méthodes cherchent à former à partir d'une population donnée une hiérarchie de partitions incluses les unes dans les autres. Les deux partitions extrêmes étant la partition discrète où chaque individu forme à lui seul une classe et la partition grossière où il n'y a qu'une seule classe formée par tous les individus. On citera parmi ces méthodes : la méthode de L.L. Cavalli-Sforza et A. Edwards, celle de S.C. Johnson, G.N. Lance et W.T. Williams et la méthode de J.H. Ward.

2.2.2. Les méthodes non hiérarchiques

Ces méthodes cherchent à diviser la population initiale en groupes disjoints, tels que, selon un critère choisi a priori, deux individus d'un même groupe ont entre eux un maximum d'affinité et deux individus de deux groupes différents ont entre eux un minimum d'affinité. Il existe dans la littérature

[2] ce qui correspond davantage à la réalité puisqu'on ne connaît pas a priori cette information

[3] pour $N = 71$, $B_{71} = 4.0811 \times 10^{74}$!

statistique une profusion de méthodes et de critères de classification automatique non hiérarchique. Mais ces méthodes s'apparentent toutes à deux familles :

- la famille des méthodes de type « centres mobiles » (ou « k-means ») telles que la méthode des nuées dynamiques de E. Diday, celle de E.W. Forgy, celle de J.B. MacQueen ou la méthode « isodata » de G.H. Ball et D.J. Hall.
- la famille des méthodes « relationnelles » telles que la méthode des partitions centrales de S. Regnier, la méthode de l'analyse relationnelle de F. Marcotorchino et P. Michaud ou la méthode des préordonnances de S. Chah.

L'algorithme de classification de RARES Text est une heuristique de la méthode de l'analyse relationnelle initialement développée par F. Marcotorchino et P. Michaud. Nous présentons cette théorie par la suite et nous expliquons brièvement en quoi, l'analyse relationnelle est une approche avantageuse par rapport aux autres méthodologies présentées ci-dessus autant du point de vue théorique que pratique.

2.3. L'analyse relationnelle et ses avantages vis à vis des autres méthodes de classification

L'analyse relationnelle créée en 1977 par F. Marcotorchino et P. Michaud s'inspire des travaux du Marquis de Condorcet, qui s'est intéressé au XVIIIème siècle au résultat collectif d'un vote à partir de votes individuels. C'est une théorie basée sur la représentation relationnelle (comparaison par paires) des différentes variables et l'optimisation sous contraintes linéaires du critère de Condorcet. L'analyse relationnelle est une théorie qui est à l'intersection de plusieurs théories :

- la théorie des graphes et l'algèbre relationnelle pour la représentation des relations et l'écriture de leurs propriétés sous forme linéaire,
- la programmation linéaire pour la résolution de problèmes multicritères dont la théorie des votes et la classification non supervisée sont des cas particuliers,
- les statistiques pour la compréhension et l'unification des mesures d'association entre variables qualitatives et des techniques classificatoires.

Parmi de nombreux champs d'applications, la classification non supervisée est un domaine dans lequel l'analyse relationnelle a permis d'avoir des avancées théoriques et pratiques conséquentes.

2.3.1. L'écriture du critère de Condorcet sous forme d'un programme linéaire

La méthodologie relationnelle s'appuie sur la définition d'une similarité $C_{ii'}$ [4] représentant le degré de ressemblance entre deux documents i et i' à partir de laquelle on construit une dissimilarité $\overline{C_{ii'}}$ représentant leur degré de dissemblance. Le critère de Condorcet à maximiser est défini a partir de ces deux définitions par la relation suivante :

$$C(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N} C_{ii'} X_{ii'} + \sum_{i=1}^{N}\sum_{i'=1}^{N} \overline{C_{ii'}}\,\overline{X_{ii'}}$$

où :

$$X_{ii'} = \begin{cases} 1 & \text{si les document } i \text{ et } i' \text{ appartiennent à une même classe} \\ 0 & \text{sinon} \end{cases}$$

et

[4] L'analyse relationnelle a été le cadre de développement de nombreuses mesures de similarités dites régularisées qui permettent de tenir compte de la structure interne des unités lexicales dans le calcul de similarité entre deux documents. Cette approche consiste à donner un poids, calculé de manière empirique, à chaque unité lexicale qui permet de mettre en avant son caractère discriminant. Nous renvoyons le lecteur intéressé aux travaux de H. Benhadda dans ce domaine.

$$\overline{X_{ii'}} = 1 - X_{i'i} = \begin{cases} 1 & \text{si les document } i \text{ et } i' \text{ n'appartiennent pas à une même classe} \\ 0 & \text{sinon} \end{cases}$$

La matrice X représente la partition finale à obtenir, elle doit respecter les contraintes d'une relation d'équivalence qui s'écrivent de manière linéaire grâce à la modélisation sous forme de comparaison par paires. Ces contraintes sont les suivantes :

- réflexivité : $X_{ii} = 1$
- symétrie : $X_{ii'} = X_{i'i}$
- transitivité : $X_{ii'} + X_{i'i''} - X_{ii''} \leq 1$

Le critère de Condorcet, en remplaçant $\overline{X_{ii'}}$ par $1 - X_{i'i}$, peut s'écrire de manière linéaire[5] de la façon suivante :

$$C(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} (C_{ii'} - \overline{C_{ii'}}) X_{ii'} + \sum_{i=1}^{N} \sum_{i'=1}^{N} \overline{C_{ii'}}$$

La classification par l'analyse relationnelle est ainsi obtenue par programmation linéaire en maximisant $C(X)$ sous les contraintes linéaires des propriétés d'une relation d'équivalence.

2.3.2. Les conditions vérifiées par le critère de Condorcet

Il a été démontré par P. Michaud et F. Marcotorchino que le critère de Condorcet, utilisé en analyse relationnelle, est un critère meilleur que les critères variationnels ou basés sur des ultramétriques puisqu'il vérifie à la fois les cinq conditions de base d'un bon critère ce qui n'est pas le cas des critères utilisés dans les autres méthodologies. Ces conditions sont résumées dans l'encadré suivant dans le cadre de la théorie des votes.

[5] Tout critère séparable en $f(C_{ii'})$ et $X_{ii'}$ relève de la même approche de linéarisation. En particulier, le cas où $f(C_{ii'}) = C_{ii'} / C_{i.}$ avec $C_{i.} = \sum_{i'} C_{ii'}$ définit le critère de Condorcet pondéré.

Condition de Non Dictature : il ne doit pas exister un juge j tel que quels que soient les candidats i et i', la préférence collective entre i et i' est celle du juge j quelles que soient les opinions des autres juges
Condition de Pareto par Paires ou d'unanimité : quels que soient les candidats i et i', si chaque juge préfère strictement i à i', alors il doit en être de même pour le choix collectif
Condition de Majorité Absolue : si la majorité des juges indique qu'un candidat i est strictement préféré à un autre candidat i' alors il doit en être de même pour le choix collectif
Condition de Neutralité Totale : l'obtention du résultat collectif doit être indépendant de l'indiçage ou de la signification des juges servant au classement ou à la classification ainsi que de l'indiçage des objets à classer ou à classifier
Condition d'Union Cohérente : si à partir de M juges de départ, on sélectionne un sous-ensemble M1 de juges donnant un résultat collectif R et qu'à partir d'un sous-ensemble disjoint M2 de juges on aboutit au même résultat collectif R, alors l'union M1 U M2 des juges doit conduire au même résultat collectif R

Tab 1. Les 5 conditions d'un bon critère d'agrégation

Si nous nous intéressons plus particulièrement à la classification des documents, les individus deviennent des documents et les juges ou variables des unités lexicales. Le critère de Condorcet permet alors de vérifier les conditions suivantes :

- condition de non dictature : cette condition signifie qu'aucune unité lexicale ne peut, par son seul profil, déterminer une classification des documents qui sera celle obtenue par la maximisation du critère de Condorcet,
- condition d'unanimité par paires de Pareto : si toute les unités lexicales sont présentes dans deux documents alors ces deux documents doivent se retrouver dans une même classe,
- condition de la majorité absolue : si sur M unités lexicales, il y a une majorité absolue d'unités lexicales qui sont partagées par deux documents alors ces deux documents doivent se retrouver dans une même classe,
- condition de neutralité totale : la classification obtenue doit être indépendante de l'ordre des documents et des unités lexicales,
- condition d'union cohérente : Si deux ensembles disjoints d'unités lexicales U_1 et U_2 donne la même partition alors la réunion des deux ensembles donnera la même partition.

2.3.3. La non fixation du nombre de classes

Contrairement aux autres méthodes, l'analyse relationnelle est une méthodologie qui respecte les données dans le sens où elle leur permet de faire émerger le nombre de classes sous-jacent sans hypothèse arbitraire.
En effet, dans la majorité des autres méthodes de classification non supervisée, la fixation d'un nombre de classes a priori est un inconvénient majeur puisque par définition fixer un nombre de classes reviendrait à chercher quelque chose que l'on connaît déjà.
On peut citer par exemple les méthodes de type « centres mobiles » qui sans ce paramétrage, aboutissent à la solution triviale où chaque individu constitue une classe et les méthodes hiérarchiques qui ont pour résultat un arbre appelé dendogramme. C'est à l'analyste de « couper » subjectivement cet arbre à un certain niveau afin d'obtenir une partition des documents avec un certain nombre de classes correspondant.

3. Présentation des caractéristiques de RARES Text

Il est extrêmement coûteux de déterminer la solution exacte du programme linéaire décrit précédemment au delà de plus d'un millier d'individus. Nous avons développé une heuristique de l'analyse relationnelle implémentée dans RARES Text dont nous présentons les caractéristiques ci-dessous. Une première heuristique de l'analyse relationnelle avait été implémentée par IBM pour ses produits Intelligent Miner for Data et for Text. Nous commençons par rappeler les caractéristiques de cette première heuristique et plus particulièrement ses inconvénients. Puis, nous montrons en quoi l'heuristique de RARES Text améliore celle d'Intelligent Miner.

3.1. Limites de l'heuristique de l'analyse relationnelle implémentée dans IBM Intelligent Miner

Intelligent Miner intègre une heuristique linéaire de l'analyse relationnelle qui permet de traiter de grandes bases de données. Cependant, ses performances en temps de traitements et en volumétrie ne sont pas accompagnées de performances en qualité. En effet, l'heuristique d'Intelligent Miner contient des limites qui sont les suivantes :

- fixation d'un nombre maximal de classes afin de préserver la caractère linéaire de l'heuristique : cet inconvénient est préjudiciable lorsque le nombre de classes « naturel[6] » du corpus est supérieur au nombre maximal de classes fixé par l'analyste. Dans ce cas, des classes très pertinentes peuvent ne pas être détectées,
- la condition de neutralité totale n'est pas vérifiée : l'ordre dans lequel sont classés les objets a un impact important sur le résultat. Il existe des artifices pour masquer ce problème tel un tri alphabétique sur le nom des objets à classifier mais on sait qu'il ne s'agit pas de solutions satisfaisantes.

3.2. Performances méthodologiques de l'heuristique de l'analyse relationnelle implémentée dans RARES Text

Nous avons développé une nouvelle heuristique qui est différente de celle implémentée dans Intelligent Miner. Nous avons réussi à dépasser les limites majeures de celle-ci.
RARES Text améliore notamment les points suivants :

- la non fixation d'un nombre maximal de classes : l'analyste n'a pas à spécifier de paramétrage à ce niveau. Le nombre de classes optimal est obtenu automatiquement, ce qui permet de détecter toutes les classes significatives[6],
- la stabilité vis à vis de l'ordre des individus : l'heuristique est par nature indépendante de l'ordre des individus et ne procède aucunement à un tri subjectif des documents avant traitement. Elle vérifie donc la condition de neutralité totale,
- la stabilité vis à vis de la duplication : si on démultiplie le corpus plusieurs fois, nous retrouvons la même solution initiale avec les individus dupliqués au sein d'une même classe,
- une très grande partie de l'heuristique est parallélisable, ce qui nous permet de bénéficier des machines multi-processeurs.

Nous rappelons aussi les avantages de RARES Text vis à vis des autres méthodologies de classification non supervisée :

- la non fixation d'un nombre de classes : ce nombre est détecté automatiquement,
- la non fixation d'un nombre d'itérations maximal : contrairement aux méthodes de type « centre mobiles » pour lesquels il faut spécifier un nombre d'itérations maximal[7],
- l'heuristique est linéaire, ce qui lui donne des temps de traitements très performants et qui ne nécessitent donc pas d'échantillonnage des documents.

[6] au sens du critère de Condorcet
[7] Plus le nombre d'itérations est grand meilleure sera la qualité de la solution obtenue

3.3. .Performances en qualité

Nous avons calculé les solutions exactes en résolvant le programme linéaire décrit en 2.2.3 sur un ensemble de 200 corpus constitué chacun de 100[8] documents et nous les avons comparées aux résultats fournis par RARES Text.
Nous avons calculé des statistiques descriptives qui illustrent les performances de l'heuristique.

Mode du taux d'erreur	0 (62 solutions exactes)
Taux d'erreur moyen	0,20%
Taux d'erreur médian	0,13%
Taux d'erreur le plus grand	0,74%

Tab 2. Statistiques relatives aux performances en qualité

Nous voyons que le taux d'erreur maximal que nous avons obtenu est strictement inférieur à 1%. Le taux d'erreur moyen sur l'ensemble des 200 tests est de 0.2%.
RARES Text améliore de l'ordre de 35 à 40 % les résultats de l'heuristique implémentée dans Intelligent Miner.

3.4. Performances en temps de traitements

RARES Text est une heuristique linéaire tant sur le nombre de documents que sur le nombre d'unités lexicales, ce qui permet donc de traiter de grands corpus dans des temps de traitements très raisonnables comme le montrent les résultats suivants.

Nom du corpus	Nombre de documents	Nombre d'unités lexicales	Temps de traitements pour la classification de base ou inférieure (en secondes)	Temps de traitements pour la hiérarchie conceptuelle (en secondes)	Temps de traitements au total (en secondes)
Newsgroup 1	20000	5000	142,68	22,97	165,65
Newsgroup 2	20000	10000	190	26,45	216,45
France 1	8366	31338	38,4	1,32	39,72
France 2	15682	90685	232,88	4,09	236,97
France 3	15682	31338	153,59	3,69	157,28
Monde 1	11831	33075	61,3	1,65	62,95
Monde 2	20233	96071	380,33	4,41	384,74

Tab 3. Exemples de temps de traitements de RARES Text

[8] Au delà de 100 individus, la résolution du programme linéaire prendrait beaucoup trop de temps sur une machine standard. Nous rappelons que la classification non supervisée est à la base un problème NP complet qui s'écrit néanmoins en analyse relationnelle de manière linéaire comme nous l'avons montré

4. L'aide à la découverte de connaissances et à la détection de signaux faibles potentiels grâce au processus d'agrégations hiérarchiques

4.1. Le processus d'agrégation hiérarchique de classes en méta-classes

Les performances de l'heuristique tant du point de vue méthodologique que technique que nous venons d'énoncer, nous ont permis de développer un processus analytique aidant à la découverte de connaissances et à la détection de signaux faibles.
En effet, nous proposons une méthode cohérente d'agrégation itérative des classes basée sur l'analyse relationnelle, donnant une hiérarchie des classes qui permet de mettre en évidence une structure arborescente de thèmes. Cette démarche s'inscrit dans le domaine de la classification conceptuelle qui vise à caractériser qualitativement et conceptuellement les résultats d'une classification en mettant davantage en valeur l'extraction de connaissances. Les approches en classification conceptuelle sont basées sur les méthodes statistiques classiques de classification que nous avons rappelées en paragraphe 1 ainsi que sur des méthodes issues de l'intelligence artificielle.
Nous illustrons dans le schéma suivant, les idées générales de l'aide à la découverte de connaissances et à la détection de signaux faibles potentiels dans RARES Text.

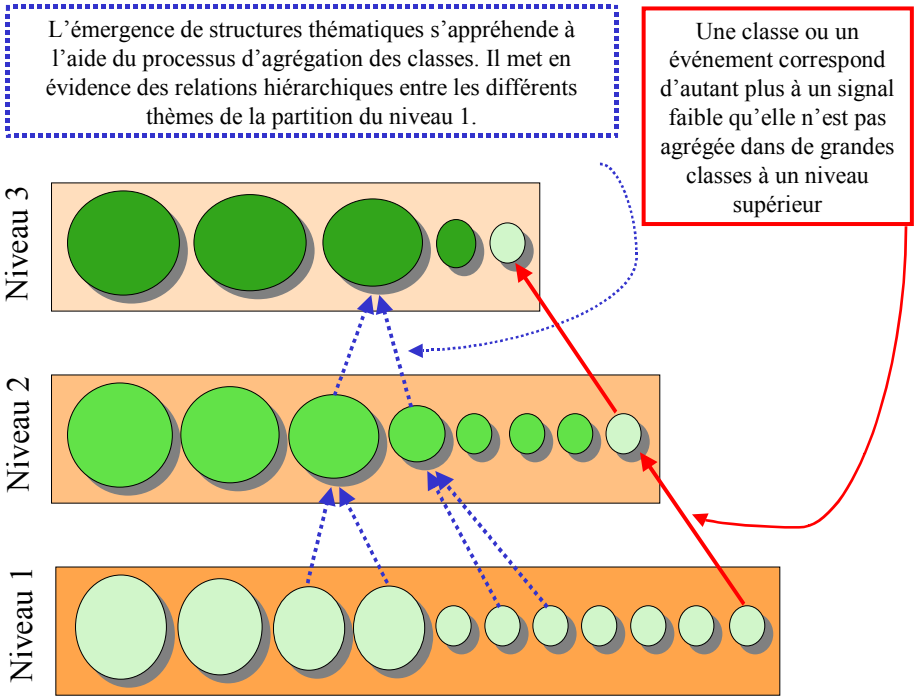

Fig 1. Processus d'agrégation hiérarchique de classes entre elles à 3 niveau, découverte de connaissances et émergence d'un signal faible potentiel

4.2. La découverte de connaissances

La découverte de connaissances est rendu possible grâce à la structuration sous forme de hiérarchie conceptuelle de l'information contenue dans un corpus. Cette hiérarchie met en évidence les relations d'inclusion des thèmes entre niveaux hiérarchiques successifs et permet une visualisation et une compréhension de l'information structurée selon divers degrés de finesse. Ainsi, le niveau le plus haut de la hiérarchie contient les méta-classes décrivant les thèmes généraux contenus dans un corpus.
Le niveau le plus bas contient, quant à lui, des classes « atomes » ou informations « atomiques » faisant émerger les thèmes les plus spécifiques du corpus.
Les niveaux intermédiaires sont les emboîtements progressifs des classes « atomes » en méta-classes et mettent en évidence de manière automatique une structure de l'information.

Des indicateurs de pertinence de la partition globale, des classes composant cette partition ainsi que les documents constituant chaque classe sont calculés pendant la classification.
RARES Text affiche pour chaque classe et chaque méta-classe de chaque niveau, les documents et les unités lexicales les plus représentatifs, ce qui permet une meilleure compréhension du résultat.
Nous illustrons le principe structuration automatique de l'information par la hiérarchie conceptuelle autour du thème sport que RARES Text a détectée lors de la classification d'un corpus de dépêches autour de l'actualité internationale :

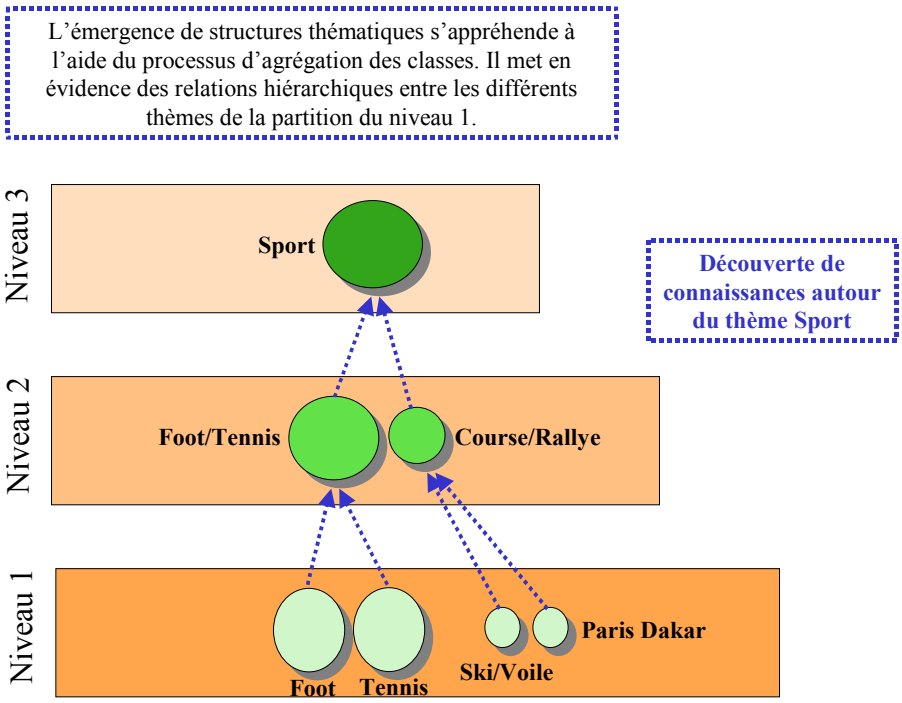

Fig 2. Illustration de la découverte de connaissances lors de la classification d'un corpus de dépêches sur l'actualité internationale d'octobre 2004 à février 2005

4.3. La détection de signaux faibles

L'analyse relationnelle et son application en classification conceptuelle permet de détecter des signaux faibles potentiels. Dans le cadre de la classification de documents, nous définirons un signal faible potentiel comme étant une information rattachée à une classe de documents de taille relativement petite, formée au niveau le plus bas de la hiérarchie conceptuelle. Ce signal, qui est détecté pour la première fois à un instant t et relativement décorrélé des autres informations contenues dans le corpus, pourra faire l'objet d'un suivi dans le temps par un analyste afin de décider de sa pertinence. Si la fréquence des documents traitant du thème relatif à ce signal augmente au cours du temps, alors le signal faible potentiel devient un signal faible effectif.

4.3.1. La détection d'un signal faible potentiel à partir de la hiérarchie conceptuelle

Grâce à la non fixation d'un nombre de classes, RARES Text est un outil qui permet la détection de signaux faibles potentiels. En effet, toute classe constituée de document apportant une information « nouvelle » et étant décrite par des unités lexicales discriminantes sera détectée.
Par ailleurs, la plausibilité qu'une information relative à une classe de documents soit un signal faible est d'autant plus grande que celle-ci s'agrége « difficilement » dans les niveaux hiérarchiques supérieurs.
En d'autre termes, une information « nouvelle » doit être relativement « orthogonale » aux autres informations contenues dans un corpus.

La décision de considérer un signal comme potentiellement faible est une tache étroitement liée a l'expertise métier de chaque domaine. RARES Text permet de proposer des candidats de signaux faibles potentiels en mettant en évidence des profils de documents plus particuliers que d'autres.
Il reviendra donc à l'analyste de traiter ces propositions en prenant les décisions adéquates.

Nous insistons sur le fait que la détection des signaux faibles potentiels est rendue possible grâce à la non fixation du nombre de classes. Le niveau inférieur qui constitue la partition de base dont les classes vont être agrégées est relativement dense en information. Cependant, elle permet de dégager naturellement des classes apportant potentiellement une information nouvelle ce qui n'aurait pas été possible par des méthodologies autres que l'analyse relationnelle.

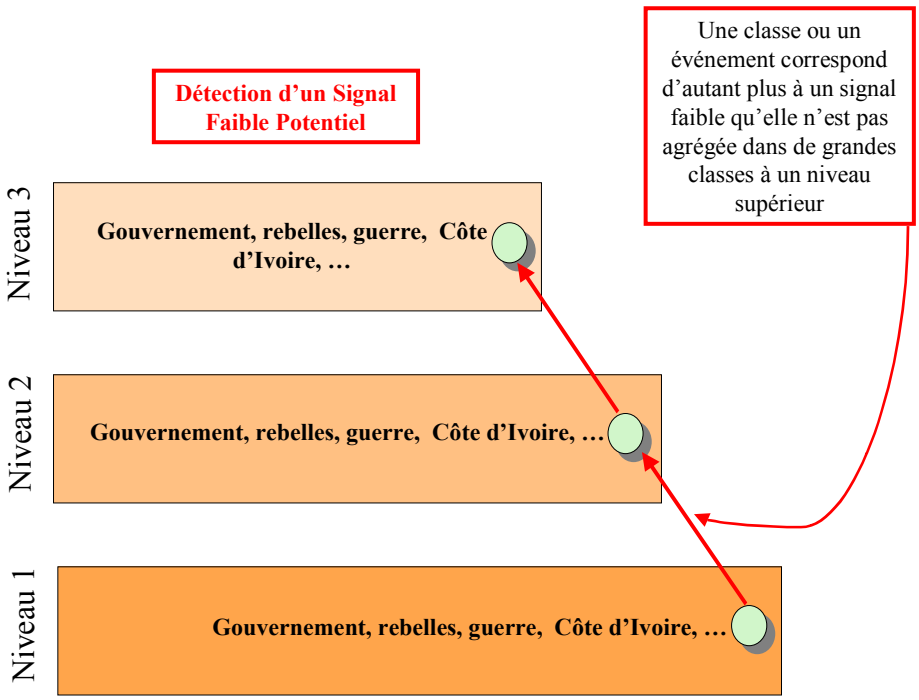

Fig 3. Illustration de la détection d'un signal faible potentiel lors de la classification d'un corpus de dépêches sur l'actualité internationale du 25 au 29 octobre 2004

4.3.2. Le suivi d'un signal faible potentiel dans le temps

Si un analyste décide qu'une classe est un signal faible potentiel, RARES Text lui permet de garder en mémoire les caractéristiques de la classe et de suivre la fréquence des documents à venir qui seront affectés à cette classe.
Nous illustrons ces propos à partir d'un exemple concernant un corpus de dépêches sur l'actualité internationale.

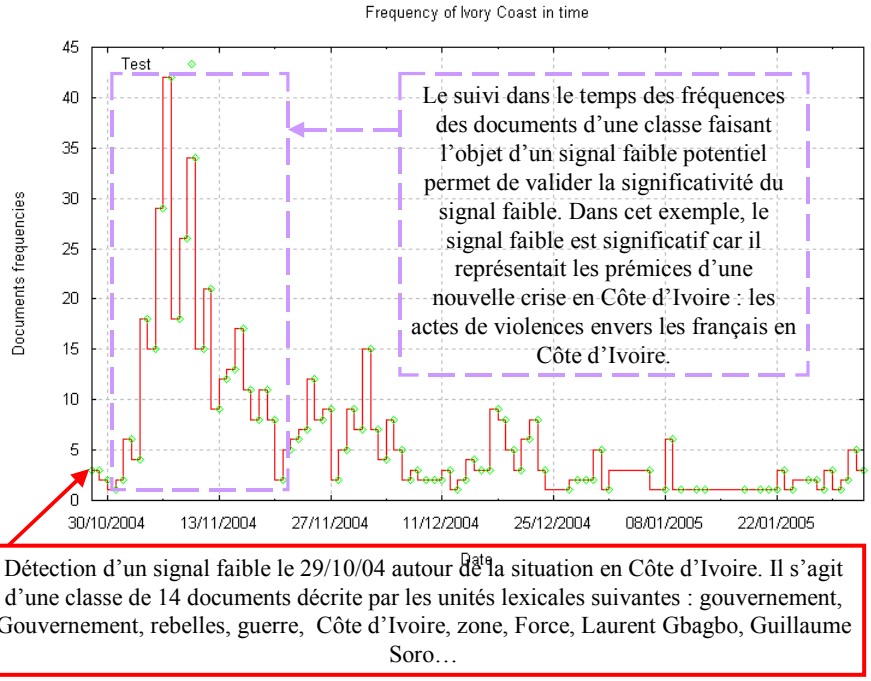

Fig 4. Illustration du suivi temporel d'un signal faible potentiel

5. Essai sur la typologie d'un signal faible à partir d'indicateurs temporel et d'extraction d'informations

Nous présentons quelques approches que nous développons dans le domaine de la typologie de signaux faibles. Ces approches reposent sur la synergie entre la linguistique computationnelle et les statistiques. Nous ne présentons pas de manière détaillée les technologies de la linguistique computationnelle. Nous précisons simplement qu'il s'agit de méthodes basées sur la définition de grammaires locales et utilisant des technologies à base d'automates à états finis permettant d'extraire des informations quantifiées d'un document telles que la gravité ou la rareté.
Nous montrons en quoi la classification non supervisée couplée aux technologies d'extraction de l'information typée permettent de détecter des signaux faibles.

Supposons que pour le corpus de documents que nous classifions, nous disposions de la date de publication t de chaque document et d'un marqueur noté g, permettant de noter la gravité d'un document sur une échelle allant de 1 à 10 par exemple. Alors, pour chaque classe issue de la classification non supervisée du corpus, nous pouvons calculer :

- la moyenne de la gravité pour la classe k (calculée sur l'ensemble des documents appartenant à la classe) notée m_g^k,
- l'écart-type de la gravité pour la classe k noté σ_g^k,
- la moyenne de la variable date[9] pour la classe k notée m_t^k,
- l'écart-type de la variable date pour la classe k noté σ_t^k,

Dans le cadre de la gravité en tant que marqueur, nous dirons qu'une classe correspond à un signal faible lorsque :

[9] Nous traitons la date de publication comme une variable numérique discrète t telle que le ou les documents les plus anciens du corpus prennent la valeur $t = 0$, le ou les documents correspondant au jour suivant prennent la valeur $t = 1$ et ainsi de suite.

- la classe représente un ensemble de documents récents (date moyenne récente) et homogènes dans le temps (écart-type plutôt faible),
- la classe a une gravité moyenne relativement faible. Dans le cas d'une gravité trop forte il ne s'agirait pas d'un signal faible par définition.

Toute classe k peut être représentée dans le plan croisant la date en abscisse et la gravité en ordonnée par une ellipse (resp. un rectangle) de centre (m_t^k, m_g^k) et d'axes (resp. de côtés) de longueur (σ_t^k, σ_g^k). La détection de signaux faibles se fait alors par visualisation du positionnement des classes vis à vis de la région que nous avons définie ci-dessus. Nous illustrons ces propos dans le schéma suivant.

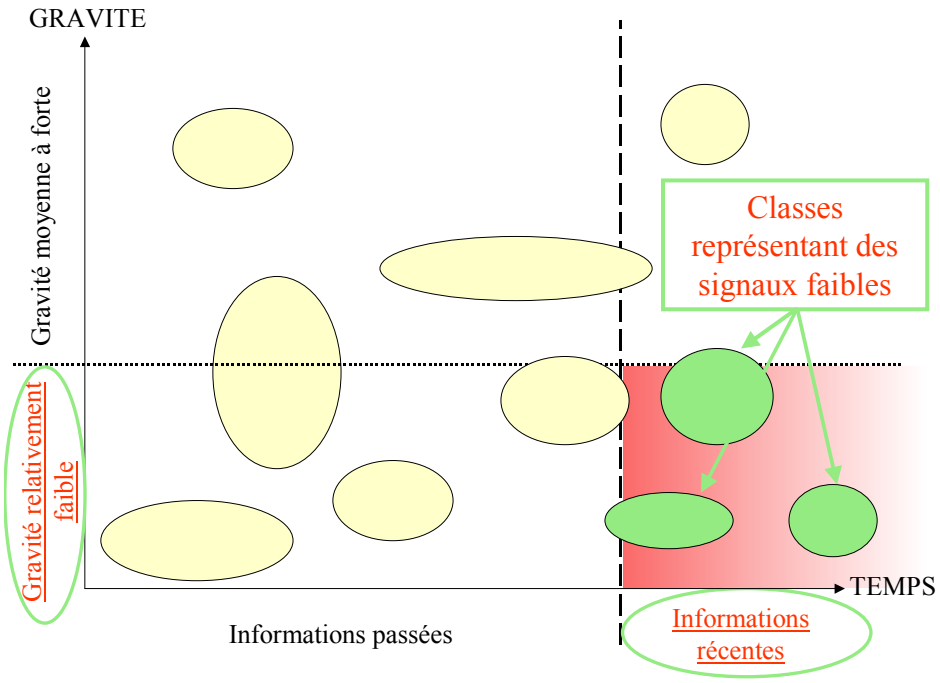

Fig 5. Typologie d'un signal faible à l'aide des marqueurs de gravité et de récence

6. Visualisation de Torgerson-Shepard des classes du niveau inférieur et de la hiérarchie conceptuelle

Pour chaque classe, nous calculons un représentant qui est le vecteur agrégé des documents appartenant à la classe. La visualisation de Torgerson-Shepard[10] permet alors de représenter les classes d'une partition dans un plan optimal au sens où l'information contenue dans le nuage de points[11] est préservée au mieux. Cette visualisation permet d'appréhender les liens entre classes puisque plus la similarité entre deux classes est grande plus leurs représentants ont une probabilité forte d'être proche dans le plan optimal. A contrario, deux classes opposées selon un ou deux axes seront très probablement de thématiques éloignées.

[10] La décomposition de Torgerson d'un produit scalaire entre deux objets permet d'avoir automatiquement le produit scalaire des deux objets centrés selon le barycentre. On obtient par ce processus une matrice carrée de mesure de similarités entre classes centrées. La diagonalisation de cette matrice permet d'avoir une décomposition du nuage de points des représentants de classe centrés dans une base orthogonale qui conserve au mieux l'information (c'est à dire les distances) contenue dans le nuage de points.
[11] Un point est ici un représentant de classes.

Par ailleurs, cette approche permet de représenter le regroupement hiérarchique des classes en méta-classes. Cette visualisation donne ainsi une représentation graphique de la découverte de connaissances.

La visualisation utilise la décomposition de Torgerson et s'inscrit dans le cadre des techniques classiques dites de « Multidimentional Scaling » qui cherchent à représenter un ensemble d'objets dans un espace euclidien à partir de matrices carrées de distances. Toutefois, les quantités calculées pour la mesure des liens entre classes et leurs agrégations hiérarchiques sont issues de l'analyse relationnelle. En d'autres termes, les fonctionnalités présentées ici reposent sur un cadre théorique cohérent. Les résultats présentés sont donc significatifs complémentaires et interprétables pour un analyste non spécialiste.
Nous illustrons la visualisation de Torgerson-Shepard dans le schéma suivant.

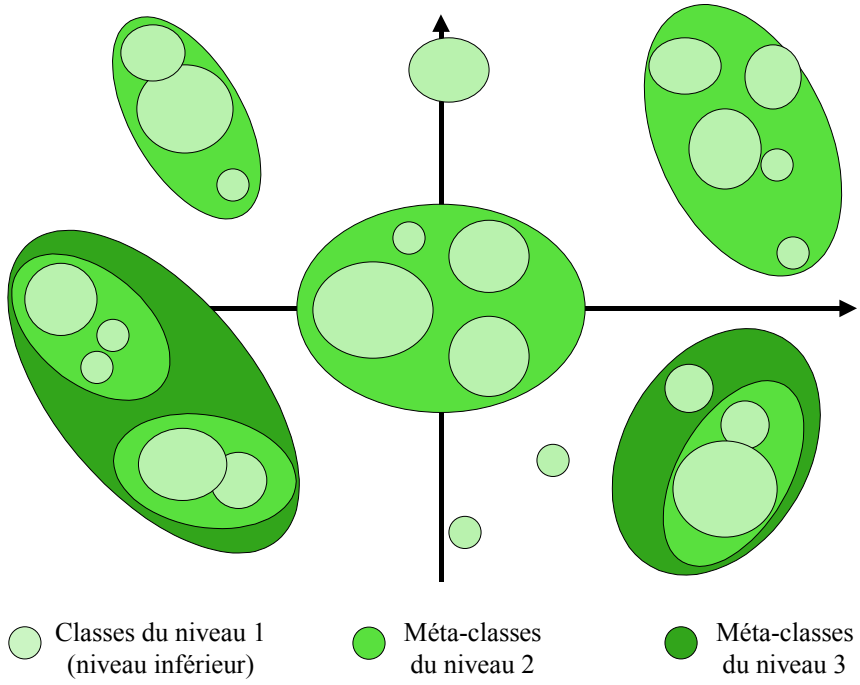

Fig 6. Visualisation graphique de Torgerson-Shepard et de la hiérarchie conceptuelle

7. Conclusion

Nous avons présenté RARES Text, un outil de classification non supervisé de documents basé sur la méthodologie de l'analyse relationnelle. RARES Text présente des caractéristiques qui en font un outil facile à utiliser, performant en terme de qualité des résultats obtenus et puissants en terme de fonctionnalités potentielles.
D'un point de vue technique, il est linéaire et permet de traiter dans des temps très raisonnables de grands corpus.
D'un point de vue méthodologique, il permet de ne pas fixer le nombre de classes ce qui autorise la détermination d'une partition naturelle et non biaisée et la détection de signaux faibles potentiels. De plus, les tests que nous avons effectués montrent que l'heuristique donne des solutions très proches des solutions exactes avec en moyenne un taux d'erreur inférieur à 1%.
D'un point de vue fonctionnel, RARES Text propose une agrégation itérative des classes de base en plusieurs niveaux hiérarchiques et une visualisation adéquate. Ceci permet à un utilisateur de mettre en évidence des hiérarchies conceptuelles et donc d'extraire des connaissances d'un ensemble de corpus.

8. Bibliographie

- Ball (G.H) et Hall (D.J.) – A clustering technique for summarizing multivariate data. *Behavorial Science*, vol 12, n°2, 1967, pp 153-155

- Bédécarrax (C.) et Huot (C.) – Développement d'indicateurs pour l'interprétation des résultats d'une analyse factorielle relationnelle. *Etude du CEMAP, IBM France,* vol MAP-05, mai 1992
- Benhadda (H.) et Marcotorchino (F.) – Introduction à la similarité régularisée en analyse relationnelle. *Revue de Statistiques Appliquées,* vol 46, n°1,1998, pp45-69
- Chah (S.) – Nouvelles techniques de codage et d'association et de classification. *Thèse de doctorat d'état de l'université Paris 6,* 1986
- Chandon (J.L.) et Pinson (S.) – Analyse Typologique : Théorie et Applications – Masson, 1980
- Decaestacker (C.) – Apprentissage en classification conceptuelle incrémentale. *Thèse de l'université Libre de Bruxelles,* 1992
- Diday (E.) – Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes. *Thèse d'état de l'université Paris 6,* 1972
- Forgy (E.W.) – Cluster analysis of multivariate data : efficiency versus interpretability of classification. *Biometrics*, vol 21, 1965, pp768-780
- Hartigan (J.) – *Clustering algorithms* – Wiley and Sons, New York, 1975
- Marcotorchino (F.) et Michaud (P.) – *Optimisation en analyse ordinale des données* – Masson, 1978
- Marcotorchino (F.) – Agrégation des similarités en classification automatique. *Thèse d'état de l'université Paris 6*, 1981
- Marcotorchino (F.) et Michaud (P.) – Agrégation des similarités en classification automatique. *Revue de Statistique Appliquée*, vol 30, n°2, 1982
- Marcotorchino (F.) – Maximal association theory as a tool of research. *In Classification as a tool of research, North Holland, Amsterdam*, 1986, pp 275-288
- Marcotorchino (F.) – La classification automatique aujourd'hui : bref aperçu historique applicatif et calculatoire. *Publications Scientifiques et Techniques d'IBM France*, n°2, Novembre 1991, pp 35-93
- Marcotorchino (F.) – L'analyse factorielle relationnelle: partie I et II. *Etude du CEMAP, IBM France,* vol MAP-03, décembre 1991
- MacQueen (J.B.) – Some methods for classification and analysis of multivariate observations. *In Berkeley Symposium on Mathematical Statistics and Probability,* éd par Le Cam (L.M.) et Neyman (J.) pp 281-297 – university of California Press
- Michalski (R.S.) et Stepp (R.E.) – Learning from observation : Conceptual clustering. *In Machine Learning : an artificial intelligence approach,* vol 1, Michalski, Carbonell, Mitchell (eds), Morgan Kaufmann, 1983
- Michaud (P.) et Marcotorchino (F.) – Modèle d'optimisation en analyse des données relationnelles. *Mathématiques et Sciences Humaines*, vol 17, n°67, 1979, pp 7-38
- Michaud (P.) – Agrégation à la majorité I : hommage à Condorcet. *Etude du Centre Scientifique IBM France,* vol F-051, 1982
- Michaud (P.) – Agrégation à la majorité II : analyse du résultat d'un vote. *Etude du Centre Scientifique IBM France,* vol F-052, 1985
- Najah Idrissi (A.) – Contribution à l'unification de critères d'association pour variables qualitatives. *Thèse de doctorat de l'université Paris 6,* 2000
- Nakache (J.P.) et Confais (J.) – *Approche pragmatique de la classification* – Technip, 2005
- Regnier (S.) – Sur quelques aspects mathématiques des problèmes de classification automatique. *ICC Bulletin*, vol 14, 1965, pp 175-191
- Saporta (G.) – *Probabilités, analyse de données et statistiques* – Technip, 1990
- Volle (M.) – *Analyse des données* – Economica, 1985, 3ème édition
- Ward (J.H.) – Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association,* vol 58, 1963, pp 236-244