Validation d'une expertise textuelle par une méthode de classification basée sur l'intensité d'implication

Jérôme David*, Fabrice Guillet*,Vincent Philippé**, Henri Briand*, Régis Gras*

*LINA - Ecole Polytechnique de l'université de Nantes
La Chantrerie - BP 50609 - 44306 Nantes cedex 3
jerome.david,fabrice.guillet,regis.gras,henri.briand@polytech.univ-nantes.fr,
http ://www.sciences.univ-nantes.fr/lina/recherche/LEC/EDC/
**PerformanSe SAS - Atlanpole - La Fleuriaye - 44470 Carquefou
vincent.philippe@performanse.fr
http ://www.performanse.fr

Résumé. Dans le cadre d'une validation d'expertise textuelle contenue dans un test de compétences comportementales informatisé, nous proposons une méthode visant à extraire des sous-ensembles de termes caractéristiques utilisés pour décrire des traits de comportements. Notre approche consiste, après une phase de traitement automatique du langage (extraction de candidats termes), à évaluer les associations possibles entre termes et étiquettes qui structurent le corpus en s'appuyant sur la théorie de l'implication statistique.

1 Introduction

Les documents sous forme de textes représentent aujourd'hui, dans notre société, des quantités d'information colossales. Afin d'accéder rapidement et de manière pertinente aux informations textuelles, des systèmes d'indexation performants permettent d'associer à un document un ensemble de caractères.

D'un autre coté, l'Extraction de Connaissances à partir de Textes (ECT) ou text-mining, vise à extraire des connaissances pertinentes, contenues dans des données textuelles, à l'aide des modèles utilisés en Extraction des Connaissances dans les Données (Kodratoff, 2000). Parmi les modèles utilisés en ECT, la découverte de règles d'associations entre termes contenus dans les textes est souvent utilisée (Maedche and Staab, 2000; Janetzko et al., 2004; Roche, 2003).

La découverte de règles d'association (Agrawal et al., 1993), consiste à trouver dans des bases de données, des tendances implicatives $a \Rightarrow b$ entre attributs booléens caractérisées par deux mesures : le support et la confiance. Parmi les indices alternatifs de qualité proposés dans la litérature (Tan et al., 2004; Guillet, 2004; Lenca et al., 2004), nous nous intéressons à la mesure d'intensité d'implication définie par R. Gras (Gras, 1979; Gras and al., 1996) et son extension entropique (Gras et al., 2001).

Cependant avant d'utiliser les techniques d'ECD, les données linguistiques doivent subir une phase de Traitement Automatique du Langage (TAL), dont le but est d'obtenir à partir d'un texte, la liste des termes qu'il contient. De nombreuses approches sont proposées : approches statistiques (Salem, 1986), approches linguistiques (David and Plante, 1990; Bourigault and Fabre, 2000; Jacquemin, 1997), ou mixtes qui combinent les deux approches précédentes (Smadja, 1993; Daille, 1994).

La démarche que nous proposons dans cet article s'inscrit à l'intersection des domaines de la recherche d'information et du text-mining. En effet, nous proposons une méthode d'étude et de validation d'une indexation par des profils psychologiques de documents traitant de bilans de compétences comportementales dans le cadre de la théorie de l'implication statistique. L'objectif de notre étude est d'associer à chaque caractère psychologique du profil, une classe de termes.

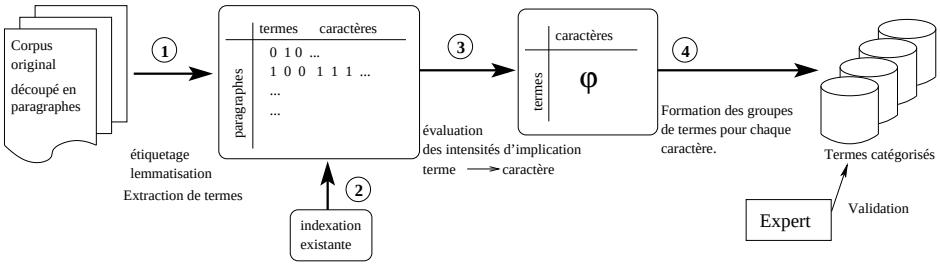

FIG. 1 – Chaîne de traitements.

Nous présentons tout d'abord, les données et la problématique à partir desquelles nous avons construit notre approche. Ensuite, nous faisons un rappel sur l'intensité d'implication. Dans la deuxième partie, nous expliquons notre démarche d'évaluation des tendances implicatives à partir desquelles nous formons des groupes de termes associés aux caractères psychologiques. Finalement, nous présentons et analysons les résultats obtenus sur la base de textes étudiée.

2 Méthodologie

2.1 Description des données analysées et de la problématique

La base textuelle indexée à partir de laquelle nous avons conçu notre méthode est extraite du logiciel PerformanSe-DIALECHO, qui est un questionnaire de personnalité informatisé largement utilisé dans le domaine de la gestion des ressources humaines. Cet outil permet, à l'issue d'un QCM composé de 70 questions, de positionner la personne évaluée sur un profil psychologique composé de 10 caractères ayant chacun 3 modalités possibles (appelés caractères) et de lui restituer un bilan de compétences sous forme textuelle. Des caractères possibles sont, par exemple, l'extraversion (EXT+), introversion (EXT-), l'anxiété (ANX+) ou encore la détente (ANX-). Un bilan de compétences est ensuite généré à partir d'une base de textes, balisée par des règles de décision (composées de conjonction de caractères), écrites par le psychologue-concepteur de l'outil.

Dans le cadre de la validation de l'expertise textuelle contenue dans ce logiciel, l'expert veut vérifier l'adéquation du vocabulaire qu'il a utilisé pour décrire un ensemble de caractères du profil psychologique.

Notre problème consiste, à associer des termes (extraits du corpus de texte) à un (ou plusieurs) caractères dont la sémantique a été définie par l'expert. Dans notre cas, ces caractères indexent également la base de textes générant le bilan de compétences.

La base de textes analysée est composée de 12805 documents. Chaque document est composé d'une conjonction de caractères et d'un paragraphe de texte.

La méthodologie que nous suivons est calquée sur le processus classique d'extraction des connaissances dans les données. En effet, une première phase de traitement et d'indexation terminologique (opération 1, figure 1), nous permet d'obtenir une représentation sous forme d'une base de données où chaque paragraphe est représenté par les termes qui le composent. Une des particularités de notre méthode, consiste à ajouter à l'indexation des documents les caractères de la conjonction décrivant le document (opération 2, figure 1). Ensuite, notre processus de fouille de données permet de former des modèles par association de termes venant de l'indexation terminologique à chaque caractère issu de l'indexation externe (opérations 3 et 4, figure 1).

2.2 Rappels sur l'analyse implicative

Les règles d'association $(a \Rightarrow b)$ sont des tendances implicatives admettant des contre-exemples $(a\, et\, \bar{b})$ dont la qualité est évaluée par les indices de support et de confiance. Nous voulons évaluer des règles qui

id_doc	conscience professionnelle	sens de la méthode	preuve de creativité	attrait de la nouveauté
d1	1	1	0	0
d2	0	0	1	1

id_doc	Extraversion	Extraversion moyenne	Rigueur	Dynamisme intellectuel
d1	0	1	1	0
d2	1	0	0	1

TAB. 1 – Extrait de la table $\mathcal{D}$ représentant les paragraphes.

peuvent avoir un faible support mais qui restent toutefois significatives. Comme le mentionne Y. Kodratoff (Kodratoff, 2001), les règles les plus pertinentes sont souvent les plus rares. La confiance présente elle aussi des défauts comme le fait de ne pas rejeter l'indépendance (Blanchard et al., 2004).

La mesure que nous avons retenue, l'intensité d'implication définie par R. Gras (Gras, 1979), quantifie l'étonnement que l'on peut avoir face à un nombre invraisemblablement petit de contre-exemples $card(A \cap \bar{B})$, où A (resp. B) sont les tuples de la relation r vérifiant a (resp. b). L'implication est évaluée sous l'hypothèse d'indépendance des variables a et b.

Pour modéliser une règle d'association, R. Gras propose, à l'instar de I.C. Lerman (Lerman, 1981), pour quantifier la similarité, de comparer le nombre de contre-exemples $card(A \cap \bar{B})$ d'une règle $a \Rightarrow b$ par rapport à la variable aléatoire $card(X \cap \bar{Y})$ où X et Y sont deux parties (choisies de manière aléatoire et indépendante) d'un ensemble E de même cardinal que la relation r étudiée.

$$\varphi(a \Rightarrow b) = 1 - Pr\left[card(X \cap \bar{Y}) \leq card(A \cap \bar{B})\right]$$

Différents modèles sont possibles pour la variable aléatoire $card(X \cap \bar{Y})$: loi Hypergéométique, loi de Poisson, loi Binomiale ou Normale. Dans le cadre de notre étude, nous nous intéressons à des cas rares par rapport la taille de la population étudiée, ainsi nous choisissons de modéliser la variable aléatoire $card(X \cap \bar{Y})$ par une loi de Poisson.

3 Regroupement des termes les plus représentatifs d'un caractère

3.1 Principes de l'étude

Nous pouvons maintenant définir la base de textes étudiée par un triplet $B = (D, T, C)$ où $D = \{d_1, .., d_m\}$ dénote l'ensemble des paragraphes, $T = \{t_1, ..., t_n\}$ l'ensemble des termes et $C = \{c_1, ..., c_y\}$ l'ensemble des caractères. Nous représentons les paragraphes par la table relationnelle $\mathcal{D}$, où chaque n-uplet représente les valeurs prises par un paragraphe d_x sur l'ensemble des attributs $A = C \cup T$. Pour un paragraphe d_x donné, un attribut a prend comme valeur 1 si l'attribut a qualifie le paragraphe d_x, 0 sinon.

L'exemple suivant (table 1) présente les valeurs prises par les paragraphes "d1" et "d2" sur l'ensemble des termes, "conscience professionnelle", "sens de la méthode", "preuve de creativité", "attrait de la nouveauté"et l'ensemble des caractères, "Extraversion", "Extraversion moyenne", "Rigueur", "Dynamisme intellectuel" :

Nous cherchons à caractériser chaque caractère par un groupe de termes représentatif. Ainsi, nous évaluons à partir de la table $\mathcal{D}$, pour chaque terme $t_i \in T$ et pour chaque caractère $c_j \in C$, l'implication $t_i \Rightarrow c_j$ signifiant : "Si un paragraphe contient le terme t_i alors ce paragraphe est destiné à un individu possédant (entre autres) le caractère c_j". Contrairement aux recherches classiques de règles d'association, nous évaluons toutes les associations binaires possibles entre deux ensembles d'attributs disjoints. Ainsi, nous

$t \Rightarrow c$	Extraversion	Extraversion moyenne	Rigueur	Dynamisme intellectuel
conscience professionnelle	0.0	0.63	0.99	0.0
sens de la méthode	0.77	0.0	0.92	0.0
preuve de créativité	0.0	0.0	0.0	0.94
attrait de la nouveauté	0.0	0.0	0.0	0.94
domaine de la communication	0.0	0.0	0.86	0.86

TAB. 2 – Extrait de la matrice $\mathcal{M}_\varphi$ d'intensités d'implication.

n'utilisons pas le support contrairement à l'algorithme Apriori (Agrawal and Srikant, 1994). Afin d'évaluer les associations, nous définissons donc, la matrice $\mathcal{M}_\varphi$ d'ordre $n \times m$ croisant les n termes et les m caractères où chaque valeur $\varphi_{ij} = \begin{cases} \varphi(t_i \Rightarrow c_j) \, si \, \varphi(t_i \Rightarrow c_j) \geq 0 \\ 0 \, sinon \end{cases}$.

La table 2 donne pour quelques termes ("conscience professionnelle", "preuve de créativité", ...) leur intensité d'implication envers les caractères ("Extraversion", "Extraversion moyenne", "Rigueur", "Dynamisme intellectuel").

Finalement, l'ensemble de termes les plus représentatifs d'un caractère c_x au seuil φ_{seuil} est défini de la manière suivante : $T_x = \{t_y \mid \varphi(t_y \Rightarrow c_x) \geq \varphi_{seuil}\}$.

Le choix du seuil φ_{seuil} est délicat car il dépend de la base de textes étudiée. Nous proposons donc fixer dans un premier temps $\varphi_{seuil} > 0,5$, car c'est le seuil à partir duquel une règle commence à être significative. Ensuite l'expert pourra l'augmenter jusqu'à satisfaire son critère de sélection : par exemple, le nombre de termes par classes.

Prenons l'exemple de la table 2 : en choisissant $\varphi_{seuil} > 0,5$, nous obtenons pour le caractère "Rigueur", la classe de termes {"conscience professionnelle", "sens de la méthode", "domaine de la communication"}. De la même manière, la classe représentative du caractère "Dynamisme intellectuel" sera constituée des termes {"preuve de créativité", "attrait de la nouveauté", "domaine de la communication"}. Nous pouvons noter que les classes ainsi formées admettent une intersection non nulle : un terme peut appartenir à plusieurs classes.

3.2 Résultats

Pour chacun des trente caractères auxquels nous voulions associer les termes les plus descriptifs, l'expert-auteur des textes a évalué la pertinence des ensembles ainsi créés. Chaque classe de termes associée à un caractère a été scindée en 2 groupes par l'expert : les termes en adéquation avec le caractère et les autres. La précision est déduite de ce classement comme la proportion de termes bien classés. Le tableau 3 donne pour quelques classes de termes associés à un caractère, la précision pour une sélection des règles ayant $\varphi_{seuil} > 0.5$.

Nous pouvons observer dans ces résultats, qu'il y a des caractères pour lesquels la précision est mauvaise (en particulier les caractères "E-", "E+" et "E0"). Les mauvais résultats sur certains ensembles de termes sont dus à la manière dont l'expert a rédigé son corpus : en effet, des caractères comme "E-", "E+" et "E0" sont très peu décrits dans les textes mais servent à nuancer la description des autres caractères étudiés. Cependant, nous avons obtenu de très bons résultats sur un bon nombre de caractères. En effet, nous avons 8 caractères pour lesquels la recherche est de bonne précision (supérieure ou égale à 90%) contre 3 caractères pour lequels les résultats sont mauvais (précision inférieure ou égale à 50%).

Classe	Précision
Rigueur (CON+)	1
Combativité (P+)	0.9
Anxiété (N+)	0.9
Dynamisme intellectuel (CLV+)	0.9
Affirmation (EST+)	0.9
Remise en cause (EST-)	0.9
Motivation de pouvoir (LED+)	0.9
Motivation de protection (LED-)	0.9
Détente (N-)	0.8
Improvisation (CON-)	0.8

Classe	Précision
Motivation d'appartenance (AFL+)	0.8
Conciliation (P-)	0.7
Motivation d'indépendance (AFL-)	0.7
Anxiété moyenne (N0)	0.6
Conformisme intellectuel (CLV-)	0.6
Introversion (E-)	0.5
Extraversion (E+)	0.4
Extraversion moyenne (E0)	0

TAB. 3 – Précisions des regroupements données par l'expert.

4 Conclusion

Nous avons présenté une approche visant à étudier et valider l'adéquation entre des termes contenus dans une base de textes et l'ensemble des caractères psychologiques indexant les paragraphes du corpus textuel. Cette méthode, divisée en deux phases (extraction et sélection des termes, formation de groupes de termes par association des termes aux caractères), permet d'obtenir pour chacun des caractères étudiés une classe de termes significatifs. L'originalité de notre approche réside dans le fait qu'elle permet de créer des rapprochements entre une indexation quelconque d'une base de textes (automatique/manuelle, ontologique...) et des termes extraits du corpus. Cela permet donc, à un expert du domaine, d'étudier et de valider la sémantique voire d'enrichir une indexation d'une base de textes.

Un prototype a été développé et appliqué au jeu de données présenté dans l'article. Les résultats sont encourageants : en effet nous avons obtenu une bonne précision moyenne des regroupements de termes, et ces derniers ont permis à l'expert d'adapter son discours en fonction du type d'individu concerné.

Actuellement, nous ne nous intéressons qu'à des caractères non structurés c'est-à-dire que l'on ne prend pas en compte les relations qui peuvent exister entre les différents caractères ou entre les termes eux-mêmes. Nous comptons donc étendre notre approche afin qu'elle puisse s'appliquer à des ontologies en prenant en compte cette dimension structurelle.

Références

Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In P., B. and S., J., editors, *Proceedings of the 1993 ACM SIGMOD ICMD*, pages 207–216.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In Bocca, J., Jarke, M., and Zaniolo, C., editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann.

Blanchard, J., Guillet, F., Gras, R., and Briand, H. (2004). Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel tic. *RNTI E-2 Extraction et gestion des connaissances*, 1 :287–298.

Bourigault, D. and Fabre, C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaires*, 25 :131–151.

Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques.* PhD thesis, University Paris 7.

David, S. and Plante, P. (1990). De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes. *ICO*, 2(3) :140–155.

Gras, R. (1979). Contribution à l'étude experimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques mathématiques. Thèse d'Etat, Université de Rennes 1.

Gras, R. and al. (1996). *L'implication statistique, une nouvelle méthode exploratoire de données.* La pensée sauvage.

Gras, R., Kuntz, P., Couturier, R., and Guillet, F. (2001). Une version entropique de l'intensité d'implication pour les corpus volumineux. *ECA Extraction et Gestion de Connaissances*, 1(1–2) :69–80.

Guillet, F. (2004). Mesure de la qualité des connaissances en ecd. In *Tuturiels de la 4ème Conf. Francophone d'extraction et gestion des connaissances*, pages 1–60, Clermond-Ferrand.

Jacquemin, C. (1997). Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes. Mémoire d'HDR, IRIN - Université de Nantes.

Janetzko, D., Cherfi, H., Kennke, R., Napoli, A., and Toussaint, Y. (2004). Knowledge-based selection of association rules for text mining. In *ECAI'04*, pages 485–489. IOS Press.

Kodratoff, Y. (2000). Datamining and textmining. In *EGC'2000*, pages 6–9.

Kodratoff, Y. (2001). On the induction of interesting rules. *Noesis*, XXVI :103–124.

Lenca, P., Meyer, P., Vaillant, B., Picouet, P., and Lallich, S. (2004). Evaluation et analyse multicritère des mesures de qualité des règles d'association. *RNTI-E-1 Mesures de qualité pour la fouille de données*, pages 219–246.

Lerman, I. (1981). *Classification et analyse ordinale des données.* Dunod, Paris.

Maedche, A. and Staab, S. (2000). Semi-automatic engineering of ontologies from text. In KSI, editor, *the 12th Internationnal Conference SEKE.*

Roche, M. (2003). L'extraction paramétrée de la terminologie du domaine. *RSTI Extraction et Gestion des Connaissances*, 17 :295–306.

Salem, A. (1986). Segments répétés et analyse statistique des données textuelles. Etude quantitative à propos du Père Duchesne de Hébert. *Histoire et Mesure*, 1(2) :5–28.

Smadja, F. (1993). Retrieving collocations from text : Xtract. *Computational linguistics*, 19 :143–177.

Tan, P., Kumar, V., and Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Inf. Syst.*, 29(4) :293–313.

Summary

In order to validate a textual base contained in a behavioural skill testing software, we suggest a methodology which can extract subsets of caracteristic terms used to describe personality traits. Our approach permits, after an automatic language processing task, to evaluate the association rules between terms and descriptors (personality traits) which structure the corpus with the help of the theory of the statistic implication. In this way, we suggest to study the inclusions between groups of terms with the cohesitive hierarchy.