

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 Chantierie - BP 50609 - 44306 Nantes cedex 3

jerome.david,fabrice.guillet,regis.gras,henri.briand@polytech.univ-nantes.fr,

**PerformanSe SAS - Atlanpole - La Fleuriaye - 44470 Carquefou

vincent.philippe@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 caractères psychologiques. Notre approche consiste, après l'extraction de termes, à évaluer les associations possibles entre termes et caractères psychologiques qui structurent le corpus en s'appuyant sur la théorie de l'implication statistique.

1 Introduction

Les documents sous forme de textes représentent des quantités d'information colossales. 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. 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).

Les règles d'association (Agrawal et al., 1993) sont 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 littérature (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 et al., 1996).

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; Jacquemin, 1997), ou mixtes qui combinent les deux approches précédentes (Smadja, 1993; Daille, 1994).

En nous inscrivant à l'intersection des domaines de la recherche d'information et du text-mining, 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 d'un profil psychologique, une classe de termes.

Validation d'une expertise textuelle basée sur l'intensité d'implication

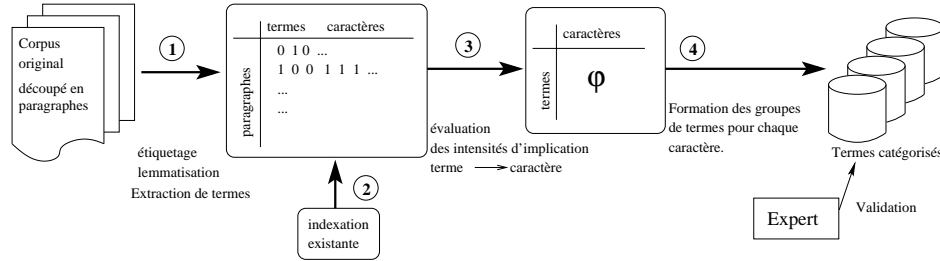


FIG. 1 – Chaîne de traitements.

Nous présentons tout d'abord, les données et la problématique à partir desquelles nous avons construit notre approche. 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 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. Cet outil permet, à l'issue d'un QCM de positionner la personne évaluée sur un profil psychologique composé de 10 dimensions à 3 modalités possibles (appelées caractères) et de lui restituer un bilan de compétences sous forme textuelle. Le bilan de compétence est 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, notre problème consiste, à associer des termes (extraits du corpus de textes) à un (ou plusieurs) caractères. 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 paragraphes, indexés par une conjonction de caractères. 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 texte est représenté par les termes qui le composent. Une des particularités de notre méthode, consiste à ajouter à la représentation des textes les caractères issus d'une autre indexation existante (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).

$t \Rightarrow c$	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. 1 – Extrait de la matrice \mathcal{M}_φ d'intensités d'implication.

3 Regroupement des termes les plus représentatifs d'un caractère

3.1 Principes de l'étude

Nous disposons d'une base de textes définie par un triplet $B = (D, T, C)$ où D dénote l'ensemble des m paragraphes, T l'ensemble des n termes et C l'ensemble des y caractères. Nous représentons les paragraphes par la table relationnelle \mathcal{D} , où chaque n -uplet représente les valeurs prises par un paragraphe d sur l'ensemble des attributs $A = C \cup T$. Pour un paragraphe d donné, un attribut a prend comme valeur 1 si l'attribut a qualifie le paragraphe d , 0 sinon.

Afin de calculer les groupes de termes associés à un seul caractère, nous évaluons l'implication $t \Rightarrow c$, entre 2 ensembles disjoints T et C , signifiant : "Si un paragraphe contient le terme t alors ce paragraphe est destiné à un individu possédant (au moins) le caractère c ". Contrairement à l'algorithme classique "A priori" (Agrawal and Srikant, 1994) nous n'utilisons pas le support. Afin d'évaluer les associations, nous calculons la matrice \mathcal{M}_φ croisant les n termes de T et les m caractères de C :

$$\varphi_{t \Rightarrow c} = \begin{cases} \varphi(t \Rightarrow c) \text{ si } \varphi(t \Rightarrow c) \geq 0 \\ 0 \text{ sinon} \end{cases}.$$

La table 1 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 au seuil φ_{seuil} est défini de la manière suivante : $T_c = \{t \mid \varphi(t \Rightarrow c) \geq \varphi_{seuil}\}$.

A partir de la table 1, 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 ne sont pas disjointes : un terme peut appartenir à plusieurs classes.

3.2 Résultats

L'expert-auteur des textes a ensuite validé l'ensemble des termes de chaque classe T_c en les scindant en 2 groupes : les termes bien classés (en adéquation avec le caractère) et les autres. Un indice de précision est ensuite déduit de ce classement comme la proportion de termes bien classés. Le tableau 2 donne pour 18 classes, leur précision (avec $\varphi_{seuil} > 0.5$).

Classe	Précision	Classe	Précision
Rigueur (CON+)	1	Motivation d'appartenance (AFL+)	0.8
Combativité (P+)	0.9	Conciliation (P-)	0.7
Anxiété (N+)	0.9	Motivation d'indépendance (AFL-)	0.7
Dynamisme intellectuel (CLV+)	0.9	Anxiété moyenne (N0)	0.6
Affirmation (EST+)	0.9	Conformisme intellectuel (CLV-)	0.6
Remise en cause (EST-)	0.9	Introversion (E-)	0.5
Motivation de pouvoir (LED+)	0.9	Extraversion (E+)	0.4
Motivation de protection (LED-)	0.9	Extraversion moyenne (E0)	0
Détente (N-)	0.8		
Improvisation (CON-)	0.8		

TAB. 2 – Précisions des regroupements données par l'expert.

Nous pouvons observer dans ces résultats, qu'il y a des caractères pour lesquels, la précision est mauvaise. 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 8 caractères pour lesquels la recherche est de bonne précision (supérieure ou égale à 90%) contre 3 caractères pour lesquels les résultats sont mauvais (précision inférieure ou égale à 50%).

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 prendre 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.
- 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 expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques mathématiques. Thèse d'Etat, Université de Rennes.
- Gras et al., R. (1996). *L'implication statistique, une nouvelle méthode exploratoire de données*. La pensée sauvage.
- 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.
- 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.
- Maedche, A. and Staab, S. (2000). Semi-automatic engineering of ontologies from text. In KSI, editor, *the 12th International 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.