Implicative statistical analysis applied to clustering of terms taken from a psychological text corpus

Jérôme David[1], Fabrice Guillet[1], Vincent Philippé[2], and Régis Gras[1]

[1] LINA - École Polytechnique de l'université de Nantes
La Chantrerie, BP 50609
44 306 NANTES Cedex 3, France
(e-mail:
jerome.david,fabrice.guillet,regis.gras@polytech.univ-nantes.fr)
[2] PerformanSe S.A.S.
Atlanpôle La Fleuriaye
44 470 CARQUEFOU, France
(e-mail: vincent.philippe@performanse.fr)

Abstract. In order to validate a textual base contained in a behavioural skill-testing software, we suggest a methodology which can extract subsets of characteristic 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) structuring the corpus with the help of the theory of statistic implication.
Keywords: data-mining, association rule, terminology, statistical implication analysis.

1 Introduction

Text-mining consists in finding knowledge structures in large text databases. To accomplish this work, text-mining uses some methods of data-mining such as association rule discovery ([Maedche and Staab, 2000], [Janetzko *et al.*, 2004], [Roche, 2003]).
Association rule discovery aims at finding implicative relations between boolean items. To evaluate these associations, the measures of support and confidence are commonly used, despite some deficiencies. In addition, many other measures are proposed ([Tan *et al.*, 2004], [Guillet, 2004], [Lenca *et al.*, 2004]). In this article, we are focusing on the implicative statistical analysis ([Gras, 1979], [Gras and others, 1996]) which offers measures such as implication intensity and entropic implication intensity ([Gras *et al.*, 2001]).
Nevertheless, we must perform some automatic language processing tasks in order to obtain a strutured list of terms representing the textual base. Many approaches are available : statistic approaches ([Salem, 1986]), linguistic approaches (([David and Plante, 1990], [Bourigault and Fabre, 2000], [Jacquemin, 1997]) and combined approaches of these two ([Smadja, 1993],

[Daille, 1994]).
In this paper, we suggest a methodology to associate each descriptor of an indexed corpus with a set of terms describing the descriptor studied. In other words, our method clusters terms with the help of other variables. First of all, we present the data, the problem and our general methodology. Then, we briefly introduce the principles of implication intensity. Next, we describe our method more precisely. Finally, we evaluate and analyse our results.

2 Methodology.

2.1 Analysed data and the problem

Our approach has been designed for a textual database extracted from a personality test, DIALECHO software, distribued by PERFORMANSE SAS company. This program is used by human resources managers. After a binary-choice questionnaire of 70 questions, this program provides a scored evaluation of 10 personality variables and a behavioural assessment report. The generation process of the report is perform in 2 steps : (1) discretisation on 3 modalities of the scored personality variables;(2) by parsing and selecting the annotated paragraphs by a conjunction of modalities named personality traits. Examples of traits: Extroversion/Introversion (discrete values : EXT+, EXT0, EXT-), Anxiety/Relaxation (ANX+, ANX0, ANX+).
Our corpus is a set of 12 805 documents. Each document is made of a paragraph (text) and a rule (conjunction of traits) as shown in figure 1. According to DIALECHO software, a document implies: "If a psychologic profile matches the rule (the conjunction) then the paragraph below will be included in the personality report". We have extracted and selected 6 977 terms from the paragraphs.

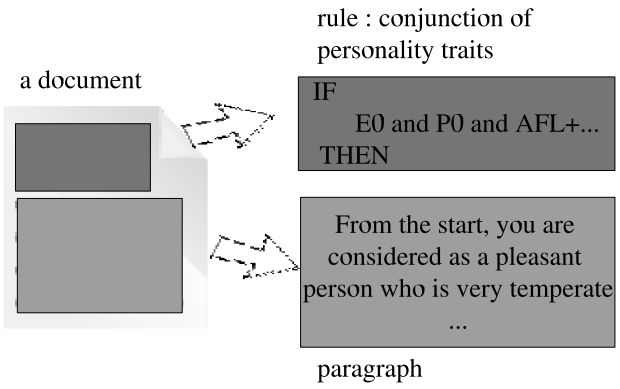

Fig. 1. The strucure of a document.

The finality of this approach is to enabled the author of the expert analysis to verify if the vocabulary used is in adequation with the personality traits described in the paragraphs. Our problem consists in finding for each item, the set of terms which best describes them.

First, we represent the paragraphs by a list of binary terms. These binary terms are noun phrases composed of two meaningful words. This terminological process (step 1, figure 2) is performed by ACABIT, an automatic term acquisition software program ([Daille, 1994]). Next, we add the personality traits set to the paragraph representation (step 2, figure 2). At this stage, a document is represented by a set of terms and personality traits. Then, we consider the set of association rules "term ⇒ trait" whose validity depends on their intensity of implication value (step 3, figure 2). For each distinct rule head (i.e. for each traits), we aim at all their bodies (terms) (step 4, figure 2). This last stage generates one cluster of terms by personality trait. Finally, the expert (author of the texts) evaluates the quality of the clusters (step 5, figure 2). According to this last stage, the expert can verify if the vocabulary used matches the personality traits.

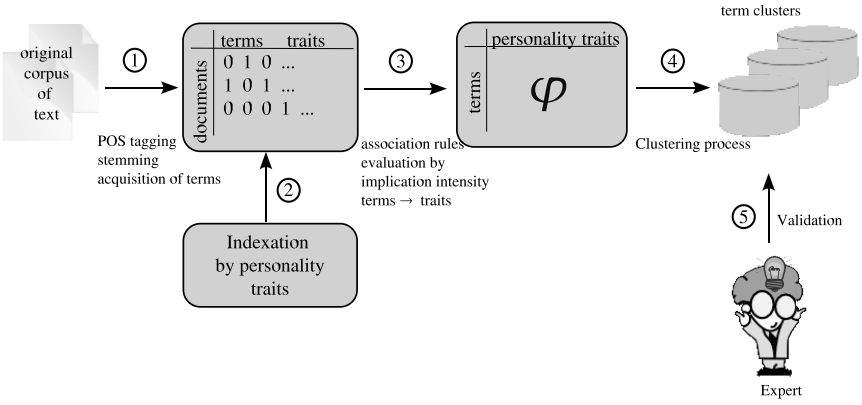

Fig. 2. Process sequence.

2.2 Rules evaluation using the implication intensity.

Association rules ([Agrawal *et al.*, 1993]) are almost like logic implications but admit some counter-examples. The quality of such rules is usually evaluated by two measures : support and confidence. Nevertheless, we intend to evaluate infrequent but interesting rules. Indeed Y. Kodratoff has said ([Kodratoff, 2001]) that "the best rules are often the least frequent". The confidence measure is not quite perfect: it cannot reject the statistical independence ([Blanchard *et al.*, 2004]).

The rule is retained for a given threshold $1-\sigma$ if $\varphi(a \Rightarrow b) \geq 1-\sigma$. Let us now consider a finite set T of n transactions described by a set I of p items. Each transaction t can be considered as an itemset so that $t \subseteq I$. A transaction t is said to contain an itemset a if $a \subseteq t$ and we denote by $A = \{t \in T; a \subseteq t\}$ the transaction set in T which contains a and by $\overline{A}$ its complementary set in T.

An association rule is an implication of the form $a \Rightarrow b$, where a and b are disjoined itemsets ($a \subset I$, $b \subset I$, and $a \cap b = \emptyset$). In practice, it is quite common to observe a few transactions which contain a and not b without contesting the general trend to have b when a is present. Therefore, with regards to the cardinal n of T but also to the cardinals n_A of A and n_B of B, the number $n_{A\cap\overline{B}} = card(A \cap \overline{B})$ of counter-examples must be taken into account to statistically accept to retain or not the rule $a \Rightarrow b$. Following the likelihood linkage analysis of Lerman [Lerman, 1981], the implication intensity expresses the unlikelihood of counter-examples $n_{A\cap\overline{B}}$ in T.

More precisely, we compare the observed number of counter-examples to a probabilistic model. Let us assume that we randomly draw two subsets X and Y in T which respectively contain n_A and n_B transactions. The complementary sets $\overline{Y}$ of Y and $\overline{B}$ of B in T have the same cardinality $n_{\overline{B}}$. In this case, $N_{X\cap\overline{Y}} = card(X \cap \overline{Y})$ is a random variable and $n_{A\cap\overline{B}}$ an observed value. The association rule $a \Rightarrow b$ is acceptable for a given threshold $1-\sigma$ if σ is greater than or equal to the probability that the number of counter-examples in the observations is greater than or equal to the number of expected counter-examples in a random drawing, i.e. if $\Pr(N_{X\cap\overline{Y}} \leq n_{A\cap\overline{B}}) \leq \sigma$.

The distribution of the random variable $N_{X\cap\overline{Y}}$ depends on the drawing mode [Gras and others, 1996]. In order to explicitly take into account the asymmetry of the relationships between itemsets, we here restrict ourselves to the Poisson distribution with $\lambda = n_A n_{\overline{B}}/n$. For cases where the approximation is justified (e.g. $\lambda > 3$), the standardized random variable $\widetilde{N}_{X\cap\overline{Y}} = (card(X \cap \overline{Y}) - \lambda)/\sqrt{\lambda}$ is approximately $N(0,1)$-distributed. The observed value of $\widetilde{N}_{X\cap\overline{Y}}$ is $\widetilde{n}_{A\cap\overline{B}} = (n_{A\cap\overline{B}} - \lambda)/\sqrt{\lambda}$.

The implication intensity of the association rule $a \Rightarrow b$ is defined by $\varphi(a \Rightarrow b) = 1 - \Pr(\widetilde{N}_{X\cap\overline{Y}} \leq \widetilde{n}_{A\cap\overline{B}})$ if $n_B \neq n$; otherwise $\varphi(a \Rightarrow b) = 0$.

The rule is retained for a given threshold $1-\sigma$ if $\varphi(a \Rightarrow b) \geq 1-\sigma$.

3 Detailed clustering process.

We choose to define the studied database by $B = (D, T, C)$ where $D = \{d_1, .., d_m\}$ is representative of the paragraph set, $T = \{t_1, ..., t_n\}$ concerns the term set and $C = \{c_1, ..., c_y\}$ express as the item set. By asserting $A = C \cup T$, the value of an item a, for a paragraph d_x, is equal to 1 if the attribute describes the document or if not to 0. The following example (table 1)

shows the values of the documents over the term set (in French): "conscience professionelle" (conscienciousness), "sens de la méthode" (rigour), "preuve de créativité" (creativity), "attrait de la nouveauté" (appeal of novelty), and the personality trait set : "Extroversion", "Medium extroversion", "Rigour", "Intellectual dynamism".

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

id_doc	Extroversion	Medium extroversion	Rigour	Intellectual dynamism
d1	0	1	1	0
d2	1	0	0	1

Table 1. Extract from the table representing the documents.

In order to build sets of terms which best describe personality traits, we evaluate for each term $t \in T$ and for each personality trait $c \in C$, the rule $t \Rightarrow c$. This rule means "if a paragraph holds the term t then this paragraph describes (at least) a person who has the personality trait c". To do this, we define the matrix $\mathcal{M}_\varphi$ of dimension $n \times m$ where rows denote terms, columns personality traits and whose values are $\psi_{t \Rightarrow c} = \begin{cases} \varphi(t \Rightarrow c) \, if \, \varphi(t \Rightarrow c) \geq 0 \\ 0 \, if not \end{cases}$.

The following example denotes the implication intensity of the French terms ("conscience professionnelle", "sens de la méthode", ...) toward the personality traits ("Extroversion", "Medium Extroversion", "Rigour", "Intellectual dynamism").

$t \Rightarrow c$	Extroversion	Medium extroversion	Rigour	Intellectual dynamism
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

Table 2. Extract from the implication intensities matrix $\mathcal{M}_\varphi$.

Finally, we define the most representative term set of a personality trait with a threshold $\varphi_{threshold}$ by the following formula :
$T_x = \{t_y \mid \varphi(t_y \Rightarrow c_x) \geq \varphi_{threshold}\}$. The choice of a threshold is not easy because it depends on the database studied. We suggest to choose, firstly, $\varphi_{threshold} = 0,5$ because a rule begin to be interesting from this threshold. After, the expert could increase this value until he/she is statisfied.

4 Results.

We have tried our method over the 30 personality traits and the expert evaluated the accuracy of each set of terms. Each set is divided into two groups by the expert (decision maker): the relevant terms for the cluster studied and the others. The accuracy value is defined as the proportion of relevant terms. The following table shows some accuracy values for groups of terms generated by our process with a threshold value $\varphi_{threshold} > 0.5$.

Class	Accuracy
Rigour (CON+)	1
Combativeness (P+)	0.9
Anxiety (N+)	0.9
Intellectual dynamism (CLV+)	0.9
Assertion (EST+)	0.9
Questioning (EST-)	0.9
Motivation of power (LED+)	0.9
Motivation of protection (LED-)	0.9
Relaxation (N-)	0.8
Improvisation (CON-)	0.8

Class	Accuracy
Motivation of belonging (AFL+)	0.8
Conciliation (P-)	0.7
Motivation of independence (AFL-)	0.7
Medium Anxiety (N0)	0.6
Intellectual conformism (CLV-)	0.6
Introversion (E-)	0.5
Extroversion (E+)	0.4
Medium extroversion (E0)	0

Table 3. Accuracy of the cluster.

Results show that some sets have bad accuracy. Indeed, these clusters describe personality traits which are not directly described in the text but their occurrence will modulate other traits. For example, the personality traits "E+", "E0", "E-" are not directly described in text but they are used to reinforce or moderate the expression of other personality traits. However, we obtain good accuracy values for most clusters. We have 8 good clusters, that is to say they have an accuracy value superior or equal to 90%. And we have only 3 bad clusters (accuracy value $< 50\%$)

5 Conclusion.

In this paper, we have presented a clustering method which matches descriptors with sets of terms based on association rules between terms and descriptors. We have designed it for a psychological corpus in order to study the adequation between terms and personality traits.
This process is divided into three steps : first, we extract a selection of relevant terms from the corpus, second, we evaluate all association rules between terms and descriptors (personality traits) with the help of implication intensity, and last, we generate sets of terms from the results obtained in the second step.
Our proposal is original in the sense that, it permits to put together terms and indexation descriptors extracted from a corpus. A prototype software program has been implemented and tested on the psychological corpus with good results.
However, we do not currently consider the relationships between descriptors or between terms. We plan to study this question in the near future in order to consider taxonomies or assimilated structures. Therefore, we intend to try our method on other corpuses.

References

[Agrawal *et al.*, 1993]R. Agrawal, T. Imielinski, and A.N. Swami. Mining association rules between sets of items in large databases. In Buneman P. and Jajodia S., editors, *Proceedings of the 1993 ACM SIGMOD ICMD*, pages 207–216, 1993.

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

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

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

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

[Gras and others, 1996]R. Gras et al. *L'implication statistique, une nouvelle méthode exploratoire de données.* La pensée sauvage, 1996.

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

[Gras, 1979]R. Gras. Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didacticques mathématiques, 1979. Thèse d'Etat, Université de Rennes.

[Guillet, 2004]F. Guillet. 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, 2004.

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

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

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

[Lenca *et al.*, 2004]P. Lenca, P. Meyer, B. Vaillant, P. Picouet, and S. Lallich. 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, 2004.

[Lerman, 1981]I.C. Lerman. *Classification et analyse ordinale des données*. Dunod, Paris, 1981.

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

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

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

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

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