LES CAHIERS DU LANCI

FROM MATHEMATICAL CLASSIFICATION TO THEMATIC ANALYSIS OF PHILOSOPHICAL TEXTS

Dominic Forest, Jean-Guy Meunier et Sophie Piron

UQÀM
Université du Québec à Montréal

No 2002-01

Le Laboratoire d'ANalyse Cognitive de l'Information (LANCI) effectue des recherches sur le traitement cognitif de l'information. La recherche fondamentale porte sur les multiples conceptions de l'information. Elle s'intéresse plus particulièrement aux modèles cognitifs de la classification et de la catégorisation, tant dans une perspective symbolique que connexionniste.
La recherches appliquée explore les technologies informatiques qui manipulent l'information. Le territoire privilégié est celui du texte.
La recherche est de nature interdisciplinaire. Elle en appelle à la philosophie, à l'informatique, à la linguistique et à la psychologie.

Publication du Laboratoire d'ANalyse Cognitive de l'Information
Directeur : Jean-Guy Meunier
Université du Québec à Montréal

Volume 3, Numéro 2002-01 – Janvier 2002

Document disponible en ligne à l'adresse suivante : www.unites.uqam.ca/lanci/

Tirage : 15 exemplaires

Dépôt legal – Bibliothèque Nationale du Canada
Dépôt legal – Bibliothèque Nationale du Québec
ISBN : 2-922916-01-4

Mise en page : L. Quevillon et D. Forest

FROM MATHEMATICAL CLASSIFICATION TO THEMATIC ANALYSIS OF PHILOSOPHICAL TEXTS

Dominic Forest, Jean-Guy Meunier et Sophie Piron
Université du Québec à Montréal

I. Introduction

Since the 1970's, research in cognitive sciences and information technologies (IT) has had an important impact on the reading and analysis of humanities texts. Various artificial intelligence classification and categorization strategies have been used to assist description and interpretation of texts. Rules and concepts are predefined and applied systematically to a text. It is a top-down analysis process. This type of strategy still highly influences what is called « *qualitative analysis* » (Glaser & Strauss, 1967), « *Grounded theory* », « *Computer assisted qualitative data analysis* » (Barry, 1998, Alexa & Zuell, 1999a) or, in more general terms, « *Computer assisted reading and analysis of text (CARAT)* ». Nevertheless, « *Qualitative research usually means the collection and analysis of unstructured textual material in order to develop concepts, categories, hypotheses, theories. Thus, most of the time during "qualitative data analysis" is spent on reading, rereading, interpreting, comparing and thinking on texts.* » (Kelle, 1997a).

On the other hand, more and more bottom-up strategies are appearing for inductive text classification and categorization. These strategies are traditionally applied to information retrieval problems, but their power can also be used in computer assisted reading and analysis of text (CARAT) for lexical analysis, automatic generation of hypertext links, knowledge extraction and thematic analysis. In this paper, we present an application of these classification and categorization technologies to one of these fields : the thematic analysis of philosophical texts.

2. *Classification*

In general terms, a classifier is an abstract machine or function that realizes some type of grouping or sorting of objects. In a set theoretical definition, classification is the projection of a partition on objects so that they are as homogenous as possible. In categorial terms, classifier can be defined as a quadruplet :

(O, X, I, G)

Where

O is the set of object $\mathbf{(o_1.... o_n)}$**,**

G is the set $\mathbf{(x_1.... x_n)}$ of features describing each object O

I is the set of type $\mathbf{(i_1.... i_n)}$

G is a discriminant function

A classifier is then a categorial operation so defined:

For each object **O,**

$\mathbf{((G(x_1.... x_n)\ _i)x_j)}$

That is, a classifier is an operation **G** that takes as input a set of objects of type **I** described by their feature **X** and delivers object of another type **I**.

The real challenge for building a good classifier is twofold : first, the classifier must be given the right descriptors for its objects and secondly it must possess the good discriminant function. This discriminant function is what allows the classifier to concretely build a class, that is : the criterion of sameness, similarity, homogeneity, equivalence etc.

Text classification can be defined as an operation, that, when applied to texts described by some of their characteristics, builds equivalent classes on them : « *Text classification is the automated grouping of textual or partially textual entities* » (Lewis & Gale, 1994).

It is important to note that classification is usually not applied to entire texts but only to fragments of texts described by their « representors » (specific words, abstracts, summaries, etc.). Secondly, classification is not an end in itself. It is a first step in a complex cognitive process of computer assisted text interpretation (Meunier, 1996). Concretely, this means that text classification must link to the dynamic process of text interpretation and hence must be appreciated according to this end.

3. Thematic analysis

Many definitions co-exist for thematic analysis. According to Popping (2000), thematic analysis can be described as the « *Identification of what, and how frequently, concepts occur in texts* ». This definition, like many others, remains very general and fuzzy. Therefore, Stone (1997) mentioned that the concept of theme « *is used in a loose, general way for analysing patterns in text* ». We agree with Stone. But, instead of describing theme analysis in terms of patterns analysis, the literature on this topic (here, we refer to studies in philosophy, literature, text analysis, etc.) shows that theme analysis lies more in the discovery and identification of the multiples relations between different themes that make a text corpus consistent and intelligible.

4. Methodology and experimentation

Our thematic analysis process is realized through 5 distinct steps : 1) The text is prepared, 2) it is then transformed into a vectorial model, 3) a classifier is applied, 4) thematic links are extracted and 5) results are interpreted.

4.1. Text preparation

To be effective, text classifiers need to define two different types of entities : the *Fragments of texts* and *the units of textual information.* Although quite simple to understand, this deconstruction of a text into FRAGS and UNIFS relies on many implicit postulates : some of them are linguistic in nature, others are mathematical.

Units of information can be either words or n-grams (chain of n characters). Choosing the units of information relies on many decisions. For instance, many possibilities are at hand : one can take into account linguistic variants of words (for instance the flexional variants such as *do, done, did*). The other possibilities are : dropping stop-words and trivial words, lemmatizing the whole corpus, aggregating or not complex lexical words such as *philosophy of mind, mind-body problem,* etc. More so, one can : retain or not all the units of information, eliminate low and high frequency words.

In this experiment, the textual data was a philosophical text : Descartes' *Discours de la méthode.* This untouched text (except for basic printing noises) contains 21 436 words. In this case the UNIFS were the lemmatized words left after eliminating hapax and functional terms. For this preparation we used an in-house program called *SATIM-Numexco.*

4.2. Text transformation into a vectorial representation

From then on, the text is transformed into a vectorial model (Salton, 1989). This procedure requires that each vector represents a segment and describes its units of information (words, n-grams, etc.). From these vectors a matrix is built (FRAGS, UNIFS). The values given to each entry of the matrix depends on the model chosen (e.g. presence, absence, fuzziness, weighting, etc.). On this textual matrix, we now have to define the core of the classification operation; that is, the *discriminan*t function. It is on this matrix that classifiers are applied.

4.3. The classifying application

In the literature of mathematical text classifiers, many types of classifiers have been proposed. All of which have their parameters; hence, fecundity and limits. The most common ones are statistically oriented (*clusterizers*, *correlators*, *factorial analysis* (Reinert, 1994), *principal component analysis* (Benzecri, 1973)). One successful implementation of these types of models has been, in the information retrieval community, the SMART system of Salton (Salton, 1989) or some variation of them such as the one found in *Latent semantic approaches* (Deerwester, 1990). Also, some more probabilistic techniques have been tested, such as *Bayesian classification*, *Nearest neighbor* (Hart, 1998), *Neural Networks classifiers* (Kohonen, 1982, Anderson, 1976), *genetic algorithm* (Holland, 1975), *Markovian fields classifiers* (Bouchaffra & Meunier, 1995b). All of these mathematical models have been applied to textual information processing.

In this experiment, we have chosen the neural ART classifier (Carpenter & Grossberg, 1991). We here explore this classifier and study its relevance for the CARAT processing, focusing on its application to thematic analysis of philosophical corpus. The purpose of this experiment is not to demonstrate its validity for texts classification, as it is often done in classical IR experiments, but instead to show its fecundity in philosophical studies.

In this ART classifier, the discriminant function measures the difference of the input segments by their weighted features (words) but through a competition amongst the segments. That is to say, the fragments that belong to a set are the winners of their weight competition. In most models, this computation consists in accepting a segment only if the impact of all the features of an input attains a certain threshold (called RHO).

The ART I classifier (Grossberg, 1988) is highly interesting. It allows plastic processing of information. When a new pattern is added, the ART system classifies it in the set of classes formed by the means of a competition process undertaken during the training phase. The essential principle of this model is based on the interaction of two levels of neurons that enter into a resonance phase such as shown in the following figure.

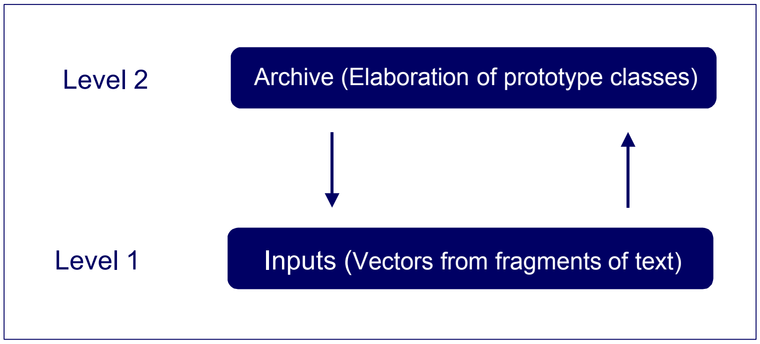
Level 2
Archive (Elaboration of prototype classes)
Level 1
Inputs (Vectors from fragments of text)

Figure 1. The ART1 resonance phase

The system receives in its first level (N1) the text fragments which are sent, after a modification of their distributed weights, to the second level of neurons (N2). This transmission implies various complex differential operations, variation of the activating forces, degradation, shunting, etc.

This second level will then possess many such modified patterns. Each of which serves as prototype for new-comers at the first level. Hence, each new input will be compared to the prototype according to a resonance criteria RHO. If the correspondence is positive then the input is entered in the class of the prototype. If not, it will be considered as an emergent prototype. The adaptability emerges by this constant modification of the levels' interconnections. And this new emerging prototype will then serve for starting up a new class. As the learning goes on, there will be a consolidation of this resonance.

At the end of this process, a set of classes of segments is produced. And from each class of segments, the lexicon is extracted. We then have for each class a specific lexical list. It is on these classes that the thematic analysis begins (see figure below).

4.4. Thematic extraction

This analysis starts by a preferred "thematic word" that can be found in one of the segments of the classes produced. In this experiment we have started with the word (or concept) « reason ». From this word, thematic relations are then identified with other thematic words belonging to other classes. In the sample results shown below, a particular analysis of Descartes' *Discours de la Méthode* starts with the word "reason" found in the class 1. This class shows the vocabulary used in the metaphysical question of the human existence as found in Descarte's philosophy. But, by the same token, we also observe that the same word (reason) operates semantically in other contexts or classes. Accordingly, the reader could then decide to discover (by exploring other classes where the word « reason » is also found) that the same concept of « reason » operates in many other different fields of Descartes' writing : such as with the mind (class 4), the knowledge (class 38), the biology (class 66).

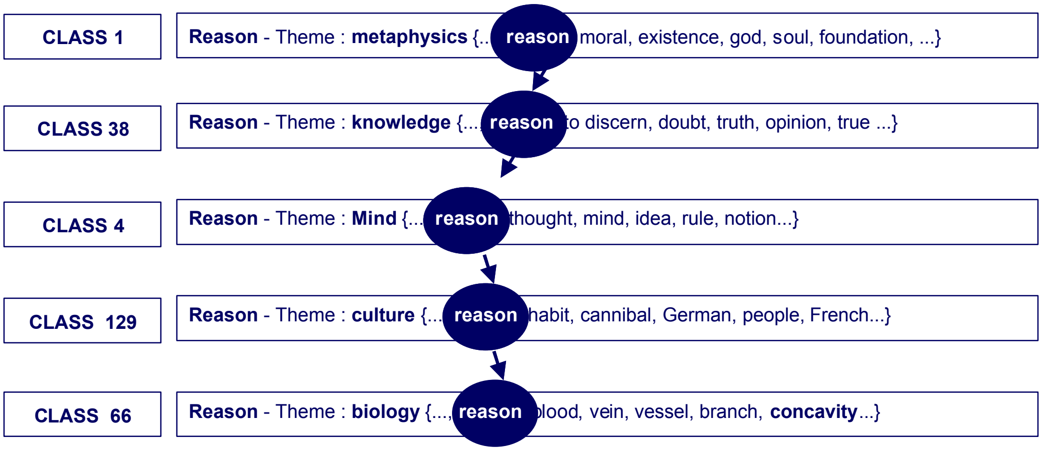
CLASS 1
Reason - Theme : metaphysics {.. reason moral, existence, god, soul, foundation, ...}
CLASS 38
Reason - Theme : knowledge {... reason to discern, doubt, truth, opinion, true ...}
CLASS 4
Reason - Theme : Mind {... reason thought, mind, idea, rule, notion...}
CLASS 129
Reason - Theme : culture {... reason habit, cannibal, German, people, French...}
CLASS 66
Reason - Theme : biology {..., reason blood, vein, vessel, branch, concavity...}

Figure 2. Thematic exploration

With this preliminary classification function, the reader can then « direct » his analysis according to a chosen theme. And from then on, he can explore other different themes found in Descartes' philosophy. For instance in the following example, he may switch to another concept inside a particular class (e.g. concavity) and start a new thematic path. This strategy is then applied to the entire text.

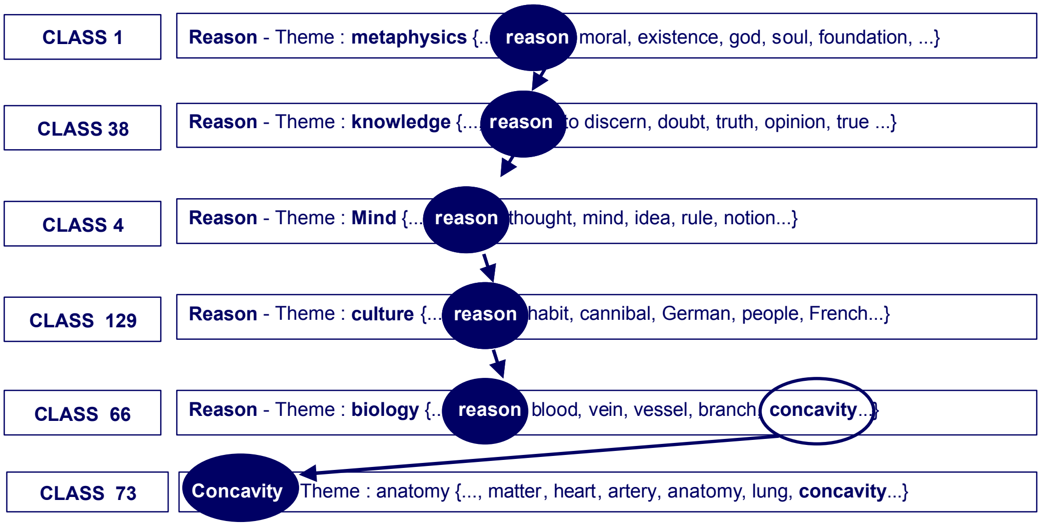
CLASS 1
Reason - Theme : metaphysics {.. reason moral, existence, god, soul, foundation, ...}
CLASS 38
Reason - Theme : knowledge {... reason to discern, doubt, truth, opinion, true ...}
CLASS 4
Reason - Theme : Mind {... reason thought, mind, idea, rule, notion...}
CLASS 129
Reason - Theme : culture {... reason habit, cannibal, German, people, French...}
CLASS 66
Reason - Theme : biology {... reason blood, vein, vessel, branch, concavity...}
CLASS 73
Concavity Theme : anatomy {..., matter, heart, artery, anatomy, lung, concavity...}

Figure 3. Other path in thematic exploration

The results have shown that many thematic links could be relevant for the reader. Also, these results, after being compared with the traditional Descartes' commentators (Rodis-Lewis, 1966 and 1985, Gueroult, 1953, Laporte, 1950, etc.), were similar with the classical interpretations of Descartes' philosophy for this particular theme.

4.5. Evaluation of classifiers for thematic analysis

From an epistemological point of view, this classification method usually relies on a benchmark produced by a set of objective experts (judges). Results are compared to these benchmark results. But we believe that reading and analysis of texts remains a subjective process. Therefore, this supposed « objective » evaluation methodology seems difficult to apply because, the discovery of the semantic content of a text is always a complex cognitive process that is guided by the reader's personal reading interests. For each reader, and mainly in philosophical texts, there is a personal interpretative path. Nevertheless, the use of such classification methods gives some « objective » orientation to the analysis. It does not impose any kind of strategy to the reader but offers him a multitude of possible paths.

One possible way to evaluate the quality of a computer thematic classification would be to compare it to various best practices or to offer the user some type of relevance feedback measures produced by HMM strategies or genetic algorithms.

Finally, we must keep in mind that these techniques using information technologies must not be seen as replacing tools for reading and analysis of text. Instead, they must be seen as « assisting tools » to help the reader's discovery and interpretation of texts.

References

Alexa, M. & C. Zuell. (1999a). *Commonalities, difference and limitations of text analysis software: The results of a review*. ZUMA arbeitsbericht, ZUMA : Mannheim.

Anderson J. R.. (1976). *Language, Memory and Tought.* New York : John Wiley & Sons.

Barry, C.A. (1998). *Choosing qualitative data analysis software : Atlas/ti and Nud*ist compared.* Sociological research online 3(3).

Benzecri, J.-P. (collab). (1973). *La Taxinomie.* Vol. I. *l'analyse des correspondances.* Dunod, Paris.

Bouchaffra, D. & Meunier, J.-G.. (1995b). *A Thematic Knowledge Extraction Modeling through a Markovian Random Field Approach.* In the *6th International DEXA 95 Conference and Workshop on Database and Expert Systems Applications*, Sept. 19-22, London, UK.

Carpenter, G. & Grossberg, G. (1991). *An Adaptive resonnance Algorithm for Rapid Category Learning and Recognition.*

Deerwester, S., Dumais. S. T., Furnas, G. Landauer. T. K. Harshman. (1990). *Indexing by latent semantic analysis. In Journal of the American Society for Information science*, pp. 391-407.

Glaser, B.G. & Strauss, A.L. (1967). *The discovery of grounded theory. Strategies for qualitative research.* Chicago : Adline.

Grossberg, S. (1988). *Neural Network and Natural Intelligence.* Cambridge : MIT Press.

Gueroult, M. (1953). *Descartes selon l'ordre des raison.* Paris : Aubier.

Hart, P. E. (1968). *The condensed nearest neighbor rule.* IEEE, Trans. on Informations theory. IT. 14 : 515.

Holland, J. (1975). *Adaptation in Natural, and Artifial Systems.* University of Michigan Press. Ann Arbor, Michigan.

Kelle, U. (1997a). *Theory Building in Qualitative Research and Computer Programs for the Management of Textual Data. In Sociological Research Online*, 2 (2)

Kohonen, T. (1982). *Clustering, taxonomy and topological Maps of Patterns. In IEEE Sixth International Conf. Pattern Recognition*, pp.114-122.

Laporte, J.M.F. (1950). *Le rationalisme des Descartes.* Paris : PUF.

Lewis, D.D. & Gale, W.A.. (1994). *A sequential algorithm for training text classifiers. In* W. Bruce Croft & C. J. van Rijsbergen (eds.). *Proceedings of the seventeenth annual international ACM-SIGIR conference on research and development in information retrieval*, pp. 3-12. Dublin : Springer-Verlag.

Meunier, J. G. (1996). *La théorie cognitive: son impact sur le traitement de l'information textuelle.* In V. Rialle et Fisette, D. (Eds.). *Penser l'Esprit, Des sciences de la cognition à une philosophie cognitive*, pp. 289-305. Grenoble : Presses UG.

Popping, R. (2000). *Computer-assisted text analysis.* London : Sage.

Reinert, M. (1994). *Quelques aspects du choix des unités d'analyse et de leur contrôle dans la méthode Alceste. In* L. L. S. Bolasco & A. Salem (eds.). *Analisi Statistica dei Dati Testuali*, vol. 1, pp.19-27. Rome : CISU.

Rodis-Lewis, G. (1966). *Descartes et le rationalisme.* Paris : PUF.

Rodis-Lewis, G. (1985). *Idées et vérités éternelles chez Descartes et ses successeurs.* Paris : J. Vrin.

Salton, G. (1989). *Automatic Text Processing.* Addison Wesley.

Stone, P. (1997). *Thematic text analysis : new agendas for analyzing text content. In* C.W. Roberts (ed.). *Text analysis for the social sciences : methods for drawing statistical inferences from texts and transcripts.* Mahwah, NJ : Lawrence Erlbaum Associates, pp. 35-54.