Draft of the paper, same title (**1998**), in: *Advances in Data Science and Classification,* A Rizzi, M. Vichy, H.-H. Bock (eds), 473-482. Springer, Berlin.

Classification problems in text analysis and information retrieval

Ludovic Lebart

Centre National de la Recherche Scientifique,
Ecole Nationale Supérieure des Télécommunications,
46, rue Barrault, 75013, Paris, France, email: lebart@eco.enst.fr

Abstract: The specific complexity of textual data sets (free answers in surveys, documentary data bases, etc.) is emphasized. Recent trends of research show that classification techniques (discrimination and unsupervised clustering as well) are widely used and have great potential in both Information Retrieval and Text Mining.

Key words: Classification; clustering; textual corpora; text mining; information retrieval.

1. The scope of multivariate analysis of texts

The amount of information available only in unstructured textual form is rapidly increasing. Classification methods play a major role in the computerized exploration of such corpora. They can contribute to process textual data sets in the three following main domains:

1.1 Producing visualizations and/or groupings of elements (free responses in marketing and socioeconomic surveys, discourses, scientific abstracts, patents, broadcast news, financial and economic reports, literary texts, etc.); looking for associations and patterns (exploratory context of Text Mining, whose ultimate aim is to extract knowledge from large bodies of unstructured textual data).

1.2 Devising decision aids for attributing a text to an author or a period; choosing a document within a database; coding information expressed in natural language.

1.3 Helping to achieve more technical or upstream contributions, such as lexical disambiguation, parsing, selection of statistical units, description of semantic graphs, speech and optical character recognition.

Note that cluster analysis has been involved in these applications ever since the beginning of these investigations (see: Jardine and van Rijsbergen, 1971; Willet, 1988).

2. Complexity and specific features of textual data

2.1 The concepts of variables and observations are more complex than those usually dealt with in most statistical applications. *Variables*, instead of being defined *a priori*, are derived from the text. Examples of (categorical) variables are the following text units: words, lemmas, segments (sequences of words appearing with a certain frequency). In the following, we use the term *word* to designate the textual unit under consideration. In lexicometry terminology, *word* is a synonym of *type*, as opposed to *token*, which designates a particular occurrence of a type.

Statistical units (or: observations, subjects, individuals, examples) are generally *documents* (described by their titles or abstracts) in documentary databases, *respondents* (described by their responses to open questions) in surveys, or segments of texts (sentences, context units, paragraphs) in literary applications. Besides the respondents, a second level of statistical units is constituted by the occurrences of words. Some statistical tests may involve counts of occurrences, whereas others deal with counts of documents or respondents. Such duality is often a source of difficulty and misunderstanding.

2.2 Two additional characteristics increase the complexity of the basic data tables: These tables are *large* (thousands of documents, thousands of words), often *sparse* (a document contains a relatively small number of words).

2.3 But the main feature of textual data sets is certainly the enormous amount of available *meta-data*. Every word is allocated several rows in a dictionary. To identify without ambiguity the lemma associated with a word may often require the help of a sophisticated syntactic analyzer. Rules of grammar, semantic networks, obviously constitute basic meta-information.

2.4 Finally, we mention that we are dealing with *sequences* of occurrences (or: strings) of items, whose order could be of importance, another non standard situation in multidimensional data analysis. Data analysts are accustomed to dealing with rectangular arrays of nominal, ordinal, or numerical variables. In textual data analysis, the basic data cannot be reduced to such an array. Let us consider the case of the responses of n individuals to an open question.

If $\{ s_1, s_2, \ldots , s_v \}$ designates a set of v different elements (the vocabulary, i. e. the set of v *different* words, in the present case), an

individual i ($i \leq n$) will be characterized by an *ordered sequence*, with *variable length* $\gamma(i)$: $\{ s_{r(i,1)}, s_{r(i,2)}, .. , s_{r(i,\gamma(i))}\}$, where $1 \leq r(i,k) \leq v$, and $1 \leq k \leq \gamma(i)$. Note that a word can appear several times in a sequence. r(i,k) is thus the index of the k^{th} word in the response of individual i. The first task of any classification method is then to compute similarities between such *sequences (with variable length) of ordered items with repetition* (see below section 4.1).

3 Clustering of observations and words

3.1 Observations (responses, documents)

The starting point is to consider each observation as described by its lexical profile, i.e. by a vector that contains the frequency of all the selected units in the text (these units could be words, at the outset). In many cases, a textual data set can lead to building a (n,v) contingency table **X** whose general entry (i,j) is the number of occurrences of word j in the text (observation) i. **X** can be easily derived from the non-rectangular array whose general entry is r(i,k) (as defined in section 2.4), but the converse is not true: the information relative to the order of the words within a response is lost in **X**. In most applications, the array r(i,k) is actually much more compact than **X**; thus, a response i containing $\gamma(i)=20$ occurrences out of a vocabulary of 2000 words corresponds to a row r(i,k) of length 20, and to a row x(i,j) of length 2000. A clustering algorithm involving computations of distances directly from the data table, such as the k-means method, can easily be reformulated using r(i,k) instead of x(i,j) in order to take advantage of the sparsity of **X**.

Note that from a given corpus, many different contingency tables can be built, according to various thresholds of frequencies for the words.

However, a usual classification algorithm applied to the rows of **X** could lead to poor or misleading results. As mentioned above, the matrix **X** could be very sparse, many pairs of rows could have no element at all in common (the computational advantage of sparsity is then pointless). Moreover, available meta-data need to be taken into account (syntactic relationships, semantic networks, external corpus and lexicons, etc.) as well as the order of occurrences within each response or text. Section 4 below will suggest some possible improvements.

3.2 Words (columns of matrix **X**)

Classification of words is rarely the final outcome of a text analysis. It is however an important intermediate step, allowing for the

definition of new statistical units, thence improving the similarities between observations (see also: section 4).
At the lower levels of a hierarchy, one can find textual co-occurrences within sentences, as well as fixed length text segments and paragraphs. The search for preferential associations is an important factor in applications involving natural language processing (see, e.g.: Lewis and Croft, 1990). It can help to solve some disambiguation problems useful, for example, in the recognition phase following optical scanning of characters.
Note that non-symmetrical measures of local associations between words (e.g.: mutual information index $I(x,y)$ resulting from the Information Theory of Shannon, as proposed by Church and Hanks (1990)) entail difficulties with classical clustering algorithms.

The main patterns observed in the first principal subspaces spanned by the first principal axes of a Singular Values Decomposition (or of a correspondence analysis) of matrix **X** are generally marked out by the higher level clusters produced by hierarchical clustering.

4 Enhancing the similarities

4.1 Taking into account the order of items

The use of additional units, such as *repeated segments* (Salem, 1984), can partially enrich the data arrays with information about order of items within texts. The repeated segment approach deals with the blind and automated detection of repeated sequences of words within a given corpus, whether or not these sequences constitute frozen phrases or expressions. The principles of a fast algorithm able to uncover such segments are given in Lebart and Salem (1997).
Direct measures of distances (such as the Levenshtein distance) have been specifically devised for strings (see, e.g.: Coggins, 1983). In this context, clustering has proved to be a crucial step when trying to extract generative grammars from a corpus of strings.

4.2 Using syntactic information

Additional variables obtained from a morpho-syntactic analyzer can be instrumental in making more meaningful the distances between lexical profiles of observations (responses or texts). The main idea is to tag the words depending on their category (nouns, verbs, preposition, etc.), and to complement the p-vector associated to each response or text with these new components of a different nature (see, e.g.: Biber, 1995; Habert and Salem, 1995, Habert et al., 1997).

4.3 Semantic relationships

The semantic information defined over the pairs of statistical units (words, lemmas) is summarized by a graph that can lead to a specific metric structure.

a)The semantic graph can be constructed from an external source of information (e.g.: a dictionary of synonyms, a thesaurus). In such a case, a preliminary lemmatization of the text must be performed. A practical way of taking into account the semantic neighbours consists in complementing the words of a given response with their semantic neighbours (provided with inferior weights). This leads to the transformation: $\mathbf{Y = X(I + \alpha M)}$, where **I** is the unit matrix, **M** the binary matrix associated with the semantic graph defined previously, and α a diffusion weight calibrating the importance given to semantic neighbourhoods. Due to this *semantic contamination*, **Y** is less sparse than **X**. The induced metric defined by the matrix: $\mathbf{Q} = \mathbf{(I+\alpha M)}^2$ leads to a new similarity index that can be used for the classification of the subjects.

b) The semantic graph can also be built up according to the associations observed in an external reference corpus, or within the corpus itself (see: Becue and Lebart, 1996). Descriptions of semantic relationships between words through self-organizing maps have been suggested by Ritter and Kohonen (1989). A hierarchical classification of words (characterized by their associates in a thesaurus), complemented with a principal axes visualization of the main nodes, produces also satisfactory descriptions of such huge graphs.

c) Another way of deriving a matrix **M** from the data themselves is to perform a hierarchical classification of words (described by their neighbours in a reference corpus) and to cut the dendrogram at a low level of the index. It can either provide a graph associated with a partition, or a more general weighted graph if the nested set of partitions (corresponding to the values of the index less than a fixed threshold) is taken into account.

5 Discriminant analysis (DA) and information retrieval

In this context, there are several outside sources of information that can be called upon to resolve classification problems: syntactic analyzers, preliminary steps toward gaining an *understanding* of the search, dictionaries or semantic networks to lemmatize and eliminate ambiguities within the search, and possibly artificial corpora that resort to experts. The major DA methods that are suited to large matrices of qualitative data are: *DA under conditional independence*,

DA by direct density estimation, DA by the method of nearest neighbours, DA on principal coordinates, neural networks.

5.1 Regularized discriminant analysis and latent semantic indexing

Regularization technique strive to make discriminant analysis possible in cases that statisticians deem to be "poorly posed" (hardly more individuals than variables) or "ill posed" (fewer individuals than variables). Correspondence analysis makes it possible to replace qualitative variables (presence or frequency of a word) with numeric variables (the values of principal coordinates), and thus to apply classical DA (linear or quadratic). It thus serves as a "bridge" between textual data (that are qualitative, and often sparse) and the usual methods of DA. But most important, a filtering of information is accomplished by dropping the last principal coordinates. This process strengthens the predictive power of the procedure as a whole (Wold, 1976). These properties are applied in Information Retrieval (IR). For instance Deerwester et al. (1990) suggest, under the name of Latent Semantic Indexing (LSI) using preliminary filtering through singular value decomposition, which is the basis of both correspondence analysis and principal components analysis. For a recent review of several filtering methods, including LSI, see: Yang (1995).

5.2 Textual units and discriminant analysis

The nature and the quality of discriminant analysis can be caused to vary depending upon the choice of the basic variables and the combination of methods:

a) Words chosen can be enhanced with counts of segments.
b) Words, and/or segments, can be selected with the help of a previously established frequency threshold.
c) The text can be lemmatized (with or without elimination of function words) and enriched with syntactic categories.
d) Only the words (segments, lemmas, etc.) that characterize the groups to be discriminated can be selected beforehand.

On the basis of the working vocabulary thus created it is possible to:

e) proceed to a preliminary singular value decomposition (or to a correspondence analysis) of the table **X** (words ∞ observations), and keep only the first axes (filtering and regularizing through SVD).
f) proceed to a preliminary cluster analysis, and work on aggregates of units (Hearst and Pedersen, 1996). In Lebart (1992), clusters are used to take into account variations of density within each category. The approaches of Salton and Mc Gill (1983), Iwayama and Tokunaga (1995), in the framework of

discriminant analysis (named in this context "Text Categorization") consist of a combination of clustering and discrimination: in a preliminary phase, clusters are built to mark out the vector space containing the observations, and to limit the number of comparisons of distances during the categorization or assignment step.

All of these alternatives imply a strategy to be developed by the user. Different strategies do exist in the framework of learning theory for using combination of methods (such as *stacking, bagging, boosting*; for a review in the specific context of IR, see: Hull *et al.*, 1996).

References

Bartell B.T., Cottrell G.W., Belew R.K. (1992). Latent semantic indexing is an optimal special case of multidimensional scaling, Proceedings of the 15th Int. *ACM-SIGIR Conf. on Res. and Dev. in Information Retrieval*, Belkin N and al. Ed., 161-167, ACM Press, New York.

Becue M., Lebart L. (1996). Clustering of texts using semantic graphs. Application to open-ended questions in surveys, *Proceedings of the IFCS 96 Symposium*, Kobe, Springer Verlag, Tokyo (in press).

Benzecri J.-P.(1977). Analyse discriminante et analyse factorielle, *Les Cahiers de l'Analyse des Donnees*, II, 4, 369-406.

Biber D. (1995). *Dimensions of register variation*. Cambridge Univ. Press, Cambridge.

Church K.W., Hanks P. (1990). Word association norms, mutual information and lexicography, *Computational Linguistics*, 16, 22-29.

Coggins J. M. (1983). Dissimilarity measures for clustering strings. in: *Time warps, string edit, and macromolecules: The theory and practice of sequence comparison*, Sankoff D. and Kruskal J. B. (eds), Addison Wesley, 311-321.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R. (1990). Indexing by latent semantic analysis, *J. of the Amer. Soc. for Information Science*, 41 (6), 391-407.

Fuhr N., Pfeifer U. (1991). Combining model-oriented and description-oriented approaches for probabilistics indexing, *Proceedings of the 14th Int. ACM Conf. on Res. and Dev. in Information Retrieval*, Bookstein A. and al., Ed, ACM Press, New York, 46-56.

Furnas G. W., Deerwester S., Dumais S.T., Landauer T.K., Harshman R. A., Streeter L.A., Lochbaum K.E. (1988). Information retrieval using a singular value decomposition model of latent semantic structure, *Proceedings of the 14th Int. ACM Conf. on Res. and Dev. in Information Retrieval*, 465-480.

Habert B., Salem A. (1995). L'utilisation de catégorisations multiples pour l'analyse quantitative de données textuelles, *Traitement Aut. des Langues*, 36, 1-2, 249-275.

Habert B., Nazarenko A., Salem A. (1997). *Les linguistiques de corpus*. Armand colin, Paris.

Hearst M. A., Pedersen J. O. (1996). Reexamining the cluster hypothesis. Scatter-Gather on retrieval results. *Proceedings of the 19th Annual International ACM SIGIR Conf. on Research and Development in Inf. Retrieval*. Zurich, 76-84.

Hull D. A.., Pedersen J. O., Schutze H. (1996). Method combination for document filtering, in: *ACM / SIGIR'96* (Frei H. P., Harman D., Schauble P., Wilkinson R., eds), Zurich, Switzerland, 279-287.

Iwayama, M., Tokunaga, T. (1995). Cluster-based text categorization: a comparison of category search strategies. in: *ACM / SIGIR'95*, (Fox E.A., Ingwersen P., Fidel R., eds), Seattle, WA, USA, 273-280.

Jardine N., van Rijsbergen C. J. (1971). The use of hierarchical clustering in information retrieval. *Information Storage and Retrieval*. 7, 217-240.

Lebart L. (1982). Exploratory analysis of large sparse matrices, with application to textual data, *COMPSTAT*, Physica Verlag, 67-76.

Lebart L. (1992). Discrimination through the regularized nearest cluster method. *COMPSTAT; Proceedings of the 10th Symposium on Computational Statistics*, Physica Verlag, Vienna, 103-118

Lebart L. (1995). Assessing and Comparing Patterns in Multivariate Analysis. in: *Data Science and its Application*. Escoufier and Hayashi (Eds), Academic Press, Tokyo, 193-204.

Lebart L., Salem A., Berry E. (1991). Recent development in the statistical processing of textual data, *Applied Stoch. Model and Data Analysis*, 7, 47-62.

Lebart L., Salem A., Berry E. (1998). *Exploring Textual Data*. Kluwer, Dordrecht.

Lewis D.D., Croft W.B. (1990). Term clustering of syntactic phrases, *Proceedings of the 13th Int. ACM Conf. on Res. and Dev. in Information Retrieval*, Vidick J.L., (ed), A.C.M.Press, New York, 385-395.

McLachlan G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New-York.

Rajman M., Rungsawang A. (1995). Textual information retrieval based on the concept of distributional semantics. In: *JADT-1995* (3rd Int. Conf. on Statist. Analysis of Textual Data), Bolasco S. et al. (eds). CISU, Roma, 237-244.

Ritter H., Kohonen T. (1989). Self Organizing Semantic Maps. *Biol. Cybern*. 61, 241-254.

Salem A. (1984). La typologie des segments répétés dans un corpus, fondée sur l'analyse d'un tableau croisant mots et textes, *Cahiers de l'Analyse des Données*, 489-500.

Salton G. (1988). *Automatic Text Processing : the Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley, New York.

Salton G., Mc Gill M.J. (1983). *Introduction to Modern Information Retrieval*, International Student Edition.

Willet P. (1988). Recent trends in hierarchical document clustering: A critical review. *Information Processing and Management*, 24 (5), 577-597.

Wold S. (1976). Pattern recognition by means of disjoint principal component models, *Pattern Recognition*, 8, 127-139.

Yang Y. (1995). Noise reduction in a statistical approach to text categorization. in: *ACM / SIGIR'95*, (Fox E.A., Ingwersen P., Fidel R., eds), Seattle, WA, USA, 256-263.