# INTENSIVE USE OF FACTORIAL CORRESPONDENCE ANALYSIS FOR TEXT MINING: APPLICATION WITH STATISTICAL EDUCATION PUBLICATIONS

Annie Morin
IRISA, Université de Rennes 1, France
amorin@irisa.fr

*Textual data are found in any survey or study and can be easily transform in frequency tables. Any method working on contingency tables can be used to process them. Besides, with the important amount of available textual data, we need to find convenient ways to process the data and to get invaluable information. It appears that the use of factorial correspondence analysis(CA) allows to get most of the information included in the data. CA produces a visual representation of the relationships between the row categories and the column categories in the same space. But there are several problems: the first one is the interpretation of the results. And even after the data processing, we still have a big amount of material and we need visualization tools to display it. In this paper, we present some methods to process the data and to get invaluable information. We also show how to use correspondence analysis in a sensible way and we give results of studies of publications dealing with statistical education.*

## INTRODUCTION

Many approaches for retrieving information from textual data depend on the literal matching of words in users' requests and those assigned to documents in a database. Generally, these approaches are concerned with the study of a lexical table which is a special 2-way contingency table. In each cell of the table, we have the occurrence of a textual unit: word, keyword, lemma. We deal with textual documents. Our goal is to get pertinent information from the data: we are doing text mining. In the past years, several methods (Hofmann, 1999; Kohonen, 1989) were proposed to process this kind of data.

The results are very promising. But there is something which is rarely mentioned: The preparation of textual data is heavy and even after processing, we are overwhelmed under a huge mass of information except if the documents we are concerned with, are monothematic, that is if there is only one topic per document.

We teach correspondence analysis and we use it to process the textual data. Actually, after processing textual data and discovering significant groups of words and/or of documents, we present the results to the experts of the field. Only these experts can evaluate the relevance of our word groupings and label the groups correctly. At this point, we need to display the results in different ways. We are not looking about finding clusters of words neither of documents. Words may have different meanings depending on the context, and may belong to different groups. Besides, a document is very often polythematic with several topics per document. Therefore, we are looking for meaningful association of words which could refer to a particular topic.

We first focus on the aspects of correspondence analysis we use to reach our goal: getting information from textual data. We explain why we prefer correspondence analysis to latent semantic analysis. We then present some display tools useful for the interpretation of the results. For illustrating the method, we first use 144 texts extracted from the *Statistics Educational Research Journal* (*SERJ*) in 2002 and 2003. Then, we study the abstracts of the *Journal of Statistics Education* (*JSE*) from 1993 to 2005. For educational purpose, the abstracts of *JSE* are very interesting.

## CORRESPONDENCE ANALYSIS

In North America, in the nineties, latent semantic indexing (LSI) and latent semantic analysis (LSA) were popularized by Deerwester, Dumais, Furnas, Landauer, and Harshman (1990) for intelligent information retrieval and for studying contingency tables. On the other hand, in France, factorial correspondence analysis (CA) is a very popular method for describing contingency tables. CA was developed 30 years ago by J. P. Benzecri in a linguistic context. The first studies with the method were performed on the tragedies of Racine. Both LSI and CA are algebraic methods whose aim is to reduce the dimension of the problem in order to make the analysis easier. Both methods use the decomposition in singular values of an ad hoc matrix

We prefer CA because the method provides indicators of the contributions by the words and by the documents to the inertia of an axis. The quality of representation of words and of documents on the various dimensions of the reduced space is also available. In CA, one of the results is the simultaneous display of the rows (documents) and of the columns (words) on a low-dimensional vector space. Generally, we have two-dimensional representations. The interpretation of an axis in CA is defined by the opposition between the most extreme points (which are very often the points with the highest contributions to inertia of the axis).

Let us have a look at Figure 1 which displays what we can obtain on the principal factorial space when our documents are monothematic. We identify A, B, C, and D as groups of words (and of documents) which define pure topics. In this case, each topic has its projection onto one axis. The interpretation is easy. There is no ambiguity among topics and we can easily identify the subject of a document.

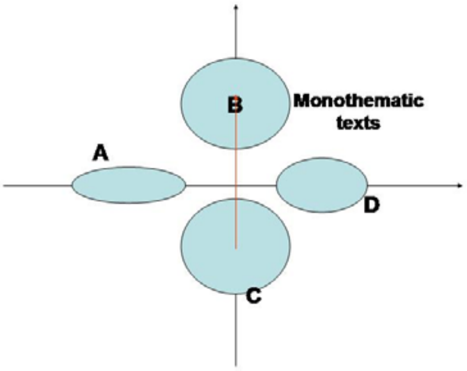


Figure 1: An ideal graph

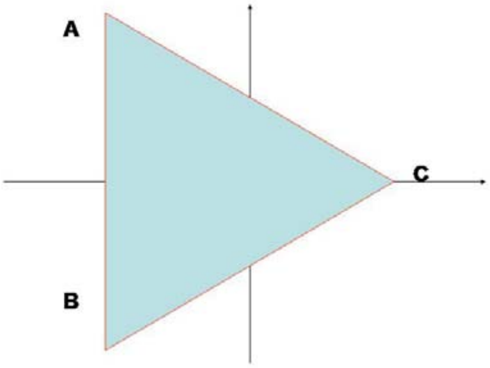


Figure 2: A frequent configuration

Figure 2 corresponds to the most frequent situation: Some topics are well represented (C for instance) on the first positive axis and is in opposition with other themes A and B. The projections of themes A and B on the left part of the first axis will be mixed up. We will get on this negative part of the axis a mixture of topics which is hard to interpret. Therefore, on each part (negative and positive) of the axes we keep, we select the words and the documents whose contributions to inertia are large, generally three times the average contribution by words or/and documents. Total inertia on an axis is equal to the corresponding eigenvalue; so the threshold is easy to compute.

M. Kerbaol calls metakeys the groups of words whose contributions are very high on one axis. Then we have two metakeys by axis, a positive one and a negative one. The metakeys at the end of this first step already can define a mixture of topics. A word can be present in several metakeys. As we keep *n* axes, a word can not be present on more than *n* metakeys among the 2*n* metakeys we obtain (one for each side of the axis).

After finding the metakeys, we can build a new contingency table crossing the metakeys. In a cell, we will have the frequency of any word present in the two related metakeys. Because of the mixture of theme in the documents, this method allows us to identify proper theme. For the preparation of data, one uses the words without no transformation. We keep all the graphical forms of a word (for instance: singular and plural). In certain situations, the plural of a word can mean another thing the singular. Therefore, the two forms may appear in different metakeys. We eliminate the “stopwords” or a list of selected words that don't bring any information in our process. After this filtering, we order the remaining words by decreasing frequency and keep the most frequent words which are present in at least α percent of the documents (α can be 2,3… or 10). At this step, some documents can be eliminated, the same with some words.

We keep the first *n* axes and get at most 2*n* metakeys. At this time, the real problem of interpretation starts and we need to work with scientific experts of the special field we are working on. We define the dimension of a word as the number of metakeys in which it appears.

Some tools make the interpretation easier: for instance, for each side of the axes, we can make a list of documents with only the words of the corresponding metakey. Thus, the expert has a summary of the contents of the documents well represented for instance on the positive side of the first axis.

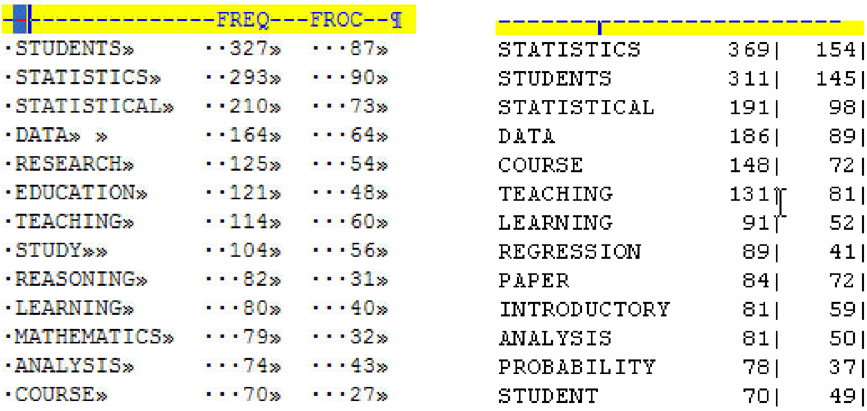
The tool *Qnomis* developed by M. Kerbaol allows us also to represent on a factorial map other criterion such as the year of publication, the center research and so on.

## PRESENTATION OF THE DATA AND VISUALIZATION OF THE RESULTS

We use texts issued from the four first numbers of the *SERJ* that is volume 1, number 1 and number 2 and volume 2, number 1 and number 2. In each volume, we select either the abstract, the introduction, or the summary of main papers, as well as recent publications and recent dissertations. We get 144 documents. Besides, we also use the abstracts issued from *JSE*: 247 abstracts. Our goal is to study the content of these documents and to try to find some relevant topics allowing an overlapping clustering of the documents. We expect also to characterize topics by a few number of associated words. We use CA and the software *BI* and *Qnomis*-3 to analyze these documents. We keep 30 axes in the CA.

After the filtering (frequency and occurrence), we get 140 documents and 464 words for *SERJ* and 224 and 576 words for *JSE*. The documents with less than 10 words were eliminated. The following table gives the first most frequent words for *SERJ* and for *JSE*.

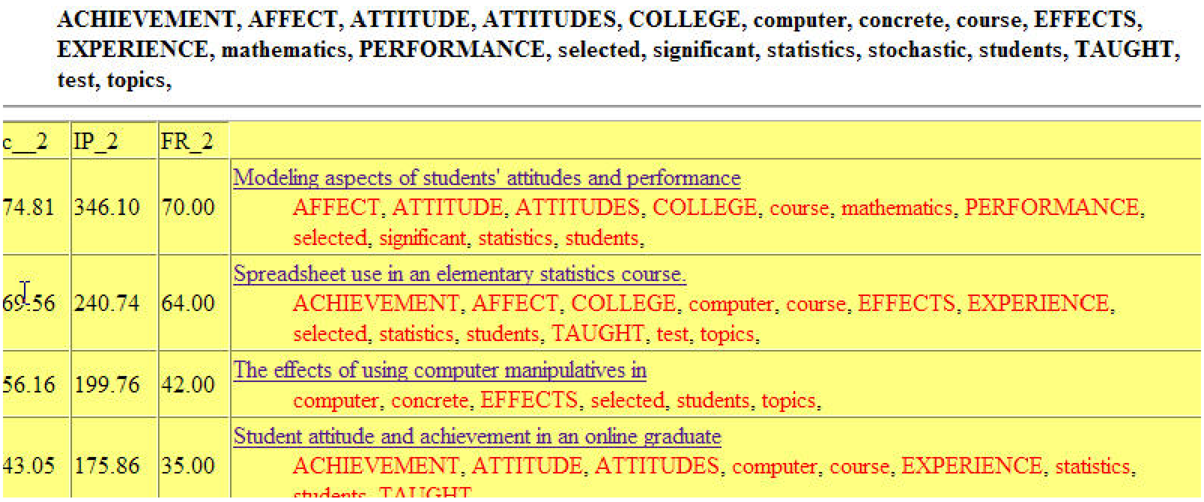
Table 1: Most frequent words in *SERJ* and *JSE*



| | FREQ | FROC |
|---|---|---|
| STUDENTS | 327 | 87 |
| STATISTICS | 293 | 90 |
| STATISTICAL | 210 | 73 |
| DATA | 164 | 64 |
| RESEARCH | 125 | 54 |
| EDUCATION | 121 | 48 |
| TEACHING | 114 | 60 |
| STUDY | 104 | 56 |
| REASONING | 82 | 31 |
| LEARNING | 80 | 40 |
| MATHEMATICS | 79 | 32 |
| ANALYSIS | 74 | 43 |
| COURSE | 70 | 27 |

| | | |
|---|---|---|
| STATISTICS | 369 | 154 |
| STUDENTS | 311 | 145 |
| STATISTICAL | 191 | 98 |
| DATA | 186 | 89 |
| COURSE | 148 | 72 |
| TEACHING | 131 | 81 |
| LEARNING | 91 | 52 |
| REGRESSION | 89 | 41 |
| PAPER | 84 | 72 |
| INTRODUCTORY | 81 | 59 |
| ANALYSIS | 81 | 50 |
| PROBABILITY | 78 | 37 |
| STUDENT | 70 | 49 |

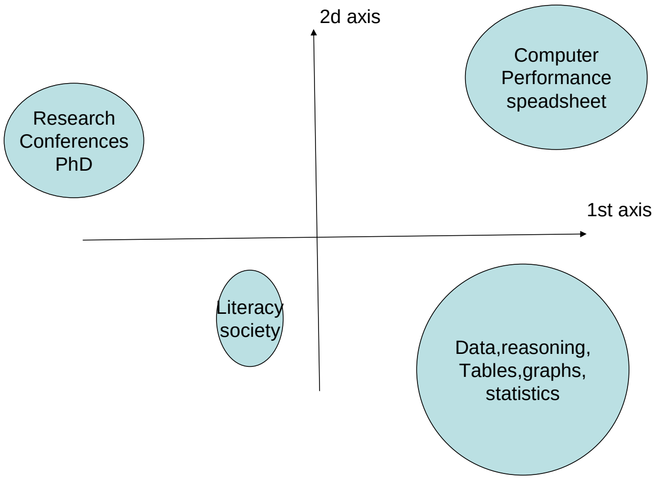
For the *SERJ* corpus, the metakey on axis 1+ contains the following words (Contribution to inertia greater than 6 times the average one) students, data and on axis 1- education, international, researchers , on axis 2+: achievement, attitudes, computer, concrete, mathematics, statistics and on axis 2- tables, literacy, thinking, statistical, data. If we reduce the contribution to 5 times the average one, we also get on the axis 1- research, statistics and to 4 times, we add the words icots, PhD doctoral, conference with two documents with a high contribution: "educating a researcher in statistics education: a personal reflection" by Pereira-Mendoza and "training future researchers in statistics education: reflections from the Spanish experience" by Batanero. As soon as we decrease the threshold, we get more documents and more words on both axis. The following results show the words whose contributions are 3 times the average contribution. In capital letters, we have the words whose contribution is the highest on this axis. The documents with the highest contributions are listed and we can click on the title of the document to get immediately the plain text.



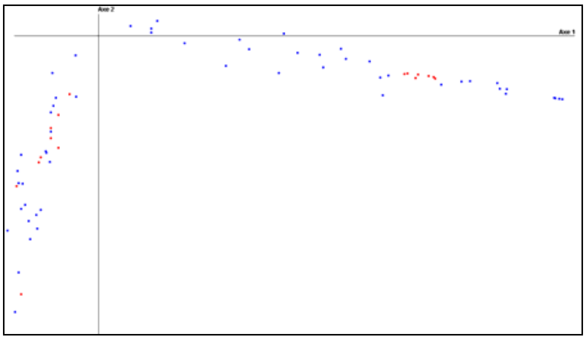
**ACHIEVEMENT, AFFECT, ATTITUDE, ATTITUDES, COLLEGE, computer, concrete, course, EFFECTS, EXPERIENCE, mathematics, PERFORMANCE, selected, significant, statistics, stochastic, students, TAUGHT, test, topics,**

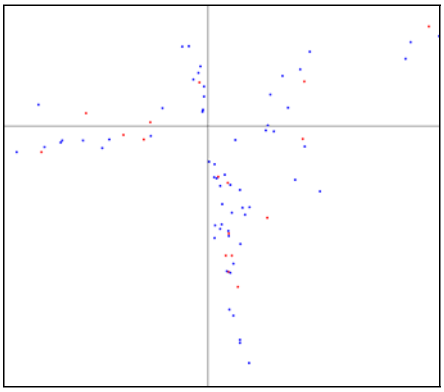
| c__2 | IP_2 | FR_2 | |
|---|---|---|---|
| 74.81 | 346.10 | 70.00 | Modeling aspects of students' attitudes and performance<br>AFFECT, ATTITUDE, ATTITUDES, COLLEGE, course, mathematics, PERFORMANCE, selected, significant, statistics, students, |
| 69.56 | 240.74 | 64.00 | Spreadsheet use in an elementary statistics course.<br>ACHIEVEMENT, AFFECT, COLLEGE, computer, course, EFFECTS, EXPERIENCE, selected, statistics, students, TAUGHT, test, topics, |
| 56.16 | 199.76 | 42.00 | The effects of using computer manipulatives in<br>computer, concrete, EFFECTS, selected, students, topics, |
| 43.05 | 175.86 | 35.00 | Student attitude and achievement in an online graduate<br>ACHIEVEMENT, ATTITUDE, ATTITUDES, computer, course, EXPERIENCE, statistics, students, TAUGHT |

The main topics on the first principal plane are the following :

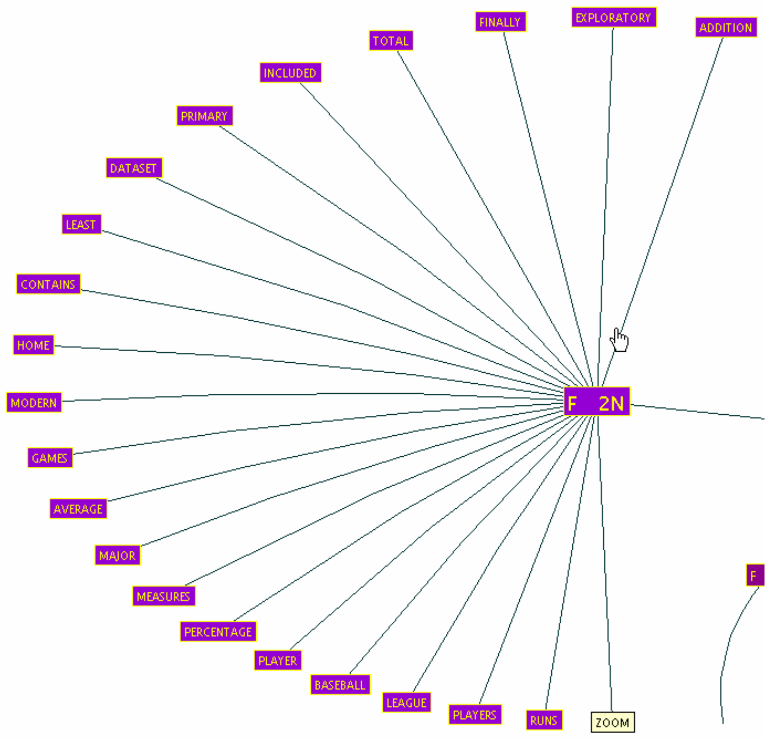


The results for *JSE* are “surprising” and interesting. The following figures display the projections (in blue words, in red abstracts) with respect to principal axes 1 and 2 (on the left) and 3 and 4 (on the right). The shapes are very characteristic.





The axis 1 is totally defined by “Teaching bits” and the axis 2 by the data (baseball, bodyweight). The axis 3 is devoted to datasets and the axis 4 to education, curriculum, learning, reasoning. The point is that in *SERJ*, there is a mixture of subject in the papers. In *JSE*, we can reorganize the contingency table crossing words and abstracts in diagonal blocks; that means that the papers are well specialized. They define clusters without great overlapping between them. It is not so frequent to find such phenomenon in real datasets. The following figure is another way of displaying the metakeys, here for the negative part of axis 2.



The correspondence analysis performed on the metakeys provides another organization of the topics. On the first principal plane, we display a group of words linked to the collaborative work, another one dealing with the applications, a third tracking the concepts and the last one concerned with the learning of randomness. Some other displays allow us to interpret the results more easily. In the end, we perform at least 4 correspondence analysis on each dataset.

## CONCLUSION

Our work is still in progress. We have to think about the interpretation of results and to help the users with displays and figures which can bring different points of view of the results. At the end of CA, the work of the statistician starts. We plan to use sequentially and automatically CA to get the greatest part of information. For textual data, CA is a very effective if the corpus is quite homogeneous although it can be used to rough out the problem.

This method was also used to select the bibliography for rare diseases. The problem with the rare diseases, one says as orphan, is that the publications with regard to them are dispersed in various fields. They are not sufficiently important to have their own magazines. First, we ask the researchers which kind of publications they read. We process the selected publications to obtain the metakeys and the vocabulary which is characteristic of the field. Some words can also characterize other medical fields. To eliminate them, one creates a database of documents with the publications concerning the rare diseases and the publications of several other great medical databases. One describes all these documents by the words selected previously, the ones of the metakeys and then carries out a CA. At the end, one preserves only the catchwords of the rare diseases.

It can be objected that there are many empirical decisions in that process, about the number of words, the number of axes and so on. The other methods have the same problem but we have to study the reliability of our choices. We plan also to study the residuals, that means the words which have not been selected at the first step. A quick study lets us think that at the first step, we recover the main research themes: it corresponds to the research strategy of an institute or a magazine and to its politics. When working on the residual words, we seem to find what is really done by the researchers, far from the fashionable topics and the magical words of the experts in communication. But as we said before, text mining is time consuming and we need helpful tools. But Correspondence analysis on text is very exciting.

ACKNOWLEDGEMENTS

Thanks to Michel Kerbaol, INSERM, LSTI, Université de Rennes 1 for his help in using the *BI* and *Qnomis* software ( michel.kerbaol@icila.fr)

REFERENCES

Benzécri, J.-P. (1973). *L'analyse des correspondances*. Paris: Dunod.

Berry, M. W. (1996). Low-rank orthogonal decompositions for information retrieval applications. *Numerical Linear Algebra with Applications*, 1(1), 1-27.

Deerwester, S., Dumais, S., Furnas, G., Landauer, K., and Harshman, R. (1990). Indexing by Latent semantic analysis. *Journal of the American Society of Information Science*, 41(6), 391-407.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Lebart L., Morineau, A., and Warwick, K. (1984). *Multivariate Descriptive Statistical Analysis*. New York: Wiley.

Lebart, L., Salem, A., and Berry, L. (1984). *Exploring Textual Data*. Dordrecht: Kluwer Academic Press.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd ACM-SIGIR International Conference on research and Development in Information retrieval*, Berkeley, CA, (pp. 50-57).

Kerbaol, M. and Bansard, J. Y. (2000). Sélection de la bibliographie des maladies rares par la technique du vocabulaire commun minimum. In M. Rajam, M. Decrauzat, J.-C. Chappelier (Eds.), *Proceedings of JADT2000: 5th Journées Internationales d'Analyse Statistique des Données Textuelles*. Lausanne: EPFL.

Kerbaol, M. and Bansard, J. Y. (2000). Pratique de l'analyse des données textuelles en bibliographie. In A. Morin, Bosc, P., Hebrail, G., and Lebart, L. (Eds.), *Bases de Données et Statistique*. Paris: Donud.

Kohonen, T. (1989). *Self Organization and Associative Memory* (3rd edition). New York: Springer-Verlag.