Intensive Use of Correspondence Analysis for Information Retrieval

Annie Morin
IRISA, Université de Rennes 1, F35042 Rennes Cedex
amorin@irisa.fr

Abstract. *With the huge 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 allows to get most of the information included in the data. Besides, 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 show how to use correspondence analysis in a sensible way and we give an application on the analysis of the internal scientific production of an important research center in France : the INRIA, the french national institute for research in computer science and control.*

Keywords. Correspondence analysis, information retrieval, textual data

1. Introduction

Many approaches for retrieving information from huge textual data depend on the literal matching of words in users'request 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, Kohonen) were proposed to process this kind of data

The results are very promising. But there is something with 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.

We use correspondence analysis to process the 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.

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. The application we are concerned with is the study of the english abstracts of internal reports of INRIA from 1989 to july 2003. We show some tools of assistance to extract information and we give some information on the display to help us in the analysis

2. Correspondence analysis

In North America, in the nineties, LSI, latent semantic indexing,and LSA, latent semantic analysis were popularized by Deerwester, Berry and Dumais (3) 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 (1) 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 [2],[4],[5],[6] 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 on the figure 1 which displays what we can obtain on the principal factorial space when our documents are monothematic. We identify A,B,C,D as groups of words (and of documents) which define pure topics . In this case, each topic has its projection on one axis. The interpretation is easy. There is no ambiguity among topics and we can easily identify the subject of a document.

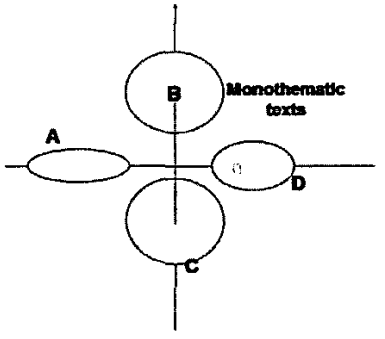

Figure 1. An ideal graph

The 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. And 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.

Kerbaol [8],[9] 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. For instance the metakey on axis 1+ for the INRIA study contains the following words (Contribution to inertia greater than 6 times the average one) :

JAVA EXECUTION
SHARED DISTRIBUTED
SEMANTICS PROGRAMS
MANAGEMENT MEMORY
SPECIFICATION PROGRAMMING
LANGUAGE PROGRAM
ARCHITECTURE SOFTWARE
PROTOCOL COMMUNICATION
ENVIRONMENT PARALLEL
DESIGN APPLICATIONS

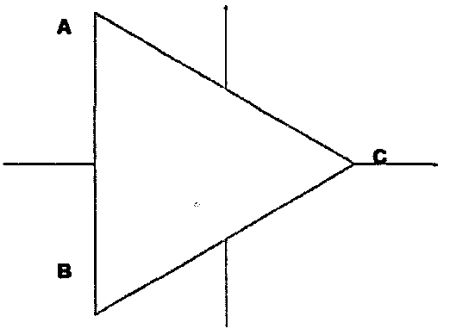

Figure 2. A frequent configuration

After finding the metakeys, we can build a new contingency table crossing the documents with the metakeys. In a cell, we will have the frequency of any word of the metakey in the document . 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 5 or 10). At this step, some documents can be eliminated , the same with some words. Most of the time, we perform a CA on a lexical table crossing the documents with approximately 850 words.

We keep the first thirty axes and get at most 60 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 words belong to the common vocabulary used in a scientific domain.

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 Kerbaol [8][9] allows us also to represent on a factorial map other criterion such as the year of publication, the center research and so on.

3. Application and visualization of the results

We use CA and the software Bi and Qnomis-3 developed by Kerbaol to analyze the internal reports of the INRIA. INRIA is an institute devoted to research in computer science, automatic and applied mathematics. It is the most important research institute in that field in France distributed on 5 centers in the country. INRIA's ambition is to be a world player, a research institute at the heart of the information society. .INRIA aims to network skills and talents from the fields of information and computer science and technology from the entire French research system.

The research is organized by topic :

1. networks and systems
2. software engineering and symbolic computing
3. human-computer interaction, images processing, data management, knowledge systems
4. simulation and optimization of complex systems

Our goal is to describe the INRIA production through the internal reports. We have 3315 reports from 1989 to July 2003, coming from the five center researchs. All topics are not equally represented. We work on the abstracts in english and keep 892 words. We are interested by the evolution of the research topics and by the productions of the different centers. We keep 30 axes in the CA.

The surprise is that the most present topics are the first one (network and system) and the last one (applied mathematics, numerical analysis which is a part of topic 4 simulation and optimization of complex system), that the evolution through the years is very soft (no change in the most productive areas). Some topics are not present in these internal reports. Therefore, we have a point of view on research with the INRIA through its internal reports.

To help us to interpret the results, we have the complete text of the abstracts. On figure 5, one represents simultaneously words and documents on the principal plane. The words are displayed on the left side and the titles of documents at the top of the screen.

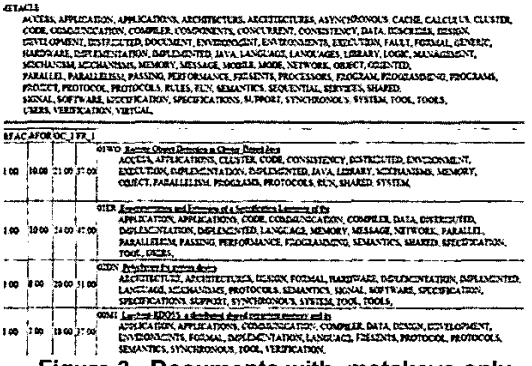

Figure 3. Documents with metakeys only

We can click on the title of the document and we get immediately the plain text.

8. Conclusion

Our work is still in progress. We plan to use sequentially and automatically CA to get the greatest part of information. We plan also to analyze each metakey and the documents well represented in its neighbourhood .

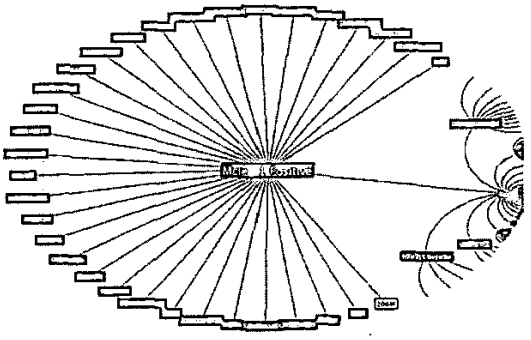

Figure 4 Representing the words in the metakeys

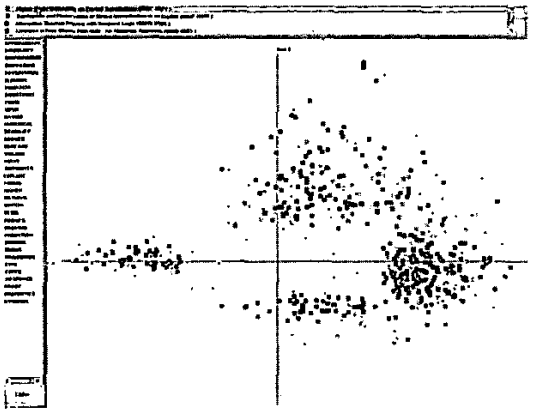

Figure 5.Projection on the principal plane.

This method was also used to select the bibliography for the rare diseases [9]. 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. Certain 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 medicaldatabases. 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 reliabilty of our choices. We plan also to study the residuals that is 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 reseach themes : it corresponds to the research strategy of an institute 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 topicss and the magical words of the experts in communication. But as we said before, text mining is time consuming and we need helpful tools.

8. Acknowledgements

Thanks to Michel Kerbaol, INSERM, LSTI, Université de Rennes 1 for his help in using the BI and Qnomis software
(michel.kerbaol@univ-rennes1.fr)

9. References

[1] Benzécri J.-P. : L'analyse des correspondances ,Dunod, Paris : 1973

[2]Berry M.W.: Low-rank orthogonal decompositions for information retrieval applications, Numerical Linear Algebra with Applications vol1 (1), 1996, 1-27

[3]Deerwester S, Dumais S,Furnas G., Landauer K, Harshman R Indexing by Ltent semantic analysis Journal of the American Society of Information Science 41(6):391-407,1990

[4] Greenacre,M.J Theory and applications of correspondence analysis, Ac. Press 1984

[5]Lebart L., Morineau A, Warwick K. Multivariate descriptive statistical analysis , J Wiley, 1984

[6] Lebart L,Salem A, Berry L. Exploring textual data , Kluwer Academic Press, 1984

[7]HofmannT. Probabilistic latent semantic indexing In Proceedings of the 22nd ACM-SIGIR International Conference on research and Development in Information retrieval 50-57, 1999

[8]Kerbaol M., Bansard J.Y. Sélection de la bibliographie des maladies rares par la technique du vocabulaire commun minimum, JADT2000 vol1 Ed Rajman, J-C Le Chapelier,ED EPFL 2000

[9]Kerbaol M., Bansard J.Y. Pratique de l'analyse des données textuelles en bibliographie, Bases de données et Statistique, Ed Morin et al., Dunod 2000

[10] Kohonen T., Self organization and associative memory Springer-Verlag, 3d edition, 1989