RIAO 94

CONFERENCE PROCEEDINGS

with presentation of prototypes and operational systems

INTELLIGENT MULTIMEDIA INFORMATION RETRIEVAL SYSTEMS AND MANAGEMENT

Rockefeller University
New York, N.Y.-U.S.A. - October 11-13, 1994

organized by

Center for the Advanced Study of Information Systems, Inc. (CASIS)

Centre de Hautes Etudes Internationales d'Informatique Documentaire (CID)

Vol. 2

R.I.A.O. = Recherche d'Information assistée par Ordinateur
(Computer aided information retrieval)

RIAO 94

NFERENCE 'RO

with presentation of prototypes
and operational systems

INTELLIGENT MULTIMEDIA INFORMATION RETRIEVAL SYSTEMS AND MANAGEMENT

Rockefeller University
New York, N.Y.-U.S.A. -October 11-13,1994

ISBN 2-905450-05-3

C.I.D. 36 bis rue Ballu 75009 PARIS FRANCE

Revealing concepts in the oral speech by a new method of indexing on some statistical criterion

Retourna C. (*), Baldit P. (*), Quoniam L. (*), M. Roux(**), Dou H. (*)

(*) CRRM - Faculte des Sciences de St Jérôme - 13397 Marseille Cedex 20

(**) Charge de Mission aux Relations Industrielles - Délégation Regionale PACA - CNRS - 31 Chemin Joseph Aiguier - 13402 Marseille Cedex 20

Abstract

The problematic of this study aims to acquire a better knowledge of phenomenon of the innovation in the firms Through semi-directing interviews conducted with some firms managers, an important volume of information has been collected under the form of oral speech

The purpose is then to develop a methodology for statistical analysis of speech to copy out in a semi-automatic mode the concepts appearing relevant in the process of innovation. So we must reduce the diversity of vocabulary encountered without important loss of information. Indeed the main difficulty of natural language analysis resides in the large diversity of vocabulary employed. A few tools and techniques used here make it possible to bring back this vocabulary to an acceptable dimension for a statistical analysis

Once this vocabulary has been reduced, some groups of words can be underlined by adapted statistical techniques These groups are reattached to concepts previously declared relevant to the innovation. These concepts, defined *a priori*, are validated or invalidated through the analysis of a set of interviews which will enable us to make a "cartography" of the innovation mechanisms in firms

Concerning the statistical analysis of interviews, it is based on some data representing the presence or the absence of words in

1- The answers to questions asked during the interview

2- The sentences contained in the whole interview

Our laboratory has been developing an algorithm of block seriation which is able to execute some regroupings by classes in a global way according to the two dimensions simultaneously analysed These classes are determined at the same time by a group of words and a group of sentences and are affected to some concepts. If there is any ambiguity on those concepts it is necessary to confirm the current list of concepts

Therefore a concept is going to regroup a set of words and sentences. Such groupings allow a new indexing to be done on some statistical criterion of collected interviews, in order to lead to a visualisation of these interviews under an hypertext mode. The user of this hypertext database will be able to move from sentence to sentence by the mean of words symbolizing the concept to be analyzed.

The final result is made of a matrix summarizing for each sentence of the interview the presence or the absence of each form of reduced vocabulary.

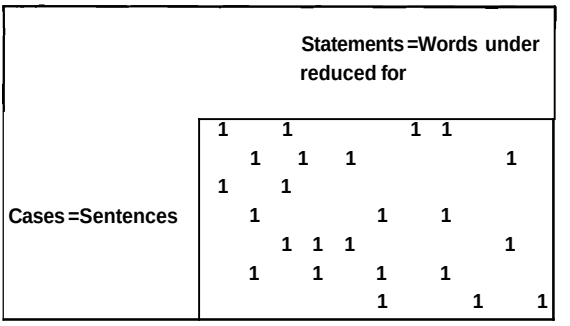

Figure 1

Statistical method

This study was carried out using a software elaborated in our laboratory, namely DATABLOC, which treats some matrices of binary data (enable/disable). By analogy with the work of F. MARCOTORCHINO and co-workers [3] in relationnal analysis of data and J.M. PROTH in operational research [4], we developped a heuristic customized to data found in science of information. This heuristic allows the treatment of a large number of data in acceptable delays on IBM PC. Although, initialy developped for bibliometric data analysis, we thought that this technic could be applied to the analysis of half closed question.

This heuristic leads to non hierarchical automatic classification of data (block seriation). The seriation proceeds simultanwusly on both spaces, the statment and the cases (columns and rows). In addition to the mathematical selection of the information by the software, it is essential to consult the interviews achieved, by reading assistance derived from the statistical analysis. Therefore a hypertext visualization of interviews should permit to scan among the sentences dealing with the same subject. A brief overview of the method used will be presented ometting the mathematical details which do not belong here.

Figure 1 representes a typical set of data before treatment.

In this specific example, the lines of this matrix correspond to the sentences and the columns to the words. The intersection of a line **i** and the column **j** is codified as follows :

1	if the word **j** is present in the sentence **i**
Blank or 0	if not

The objective is to obtain some blocks of high density of 1 on the diagonal by permuting the lines and the columns of the original matrix as illustrated figure 2.

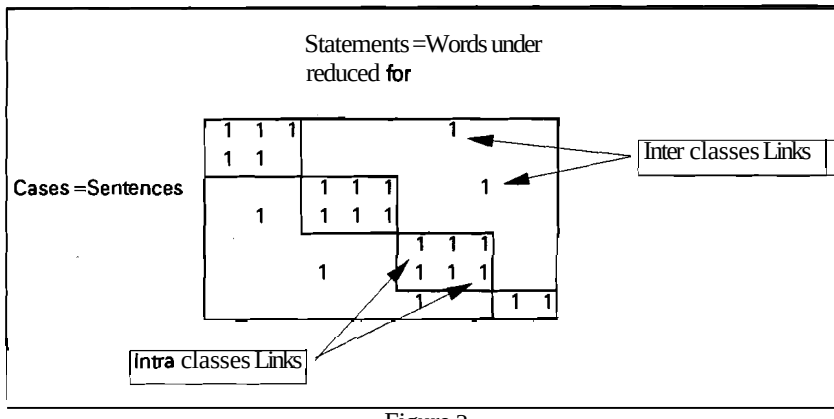

Figure 2

Each block then represents a cluster of sentences composed of common words depending on the data processed in these gathering. Depending on the data processed, these groups could have different meanings. The software enables the isolation of the words present in most of the sentences (words present everywhere). This separation is carried out automatically according to the statistical distribution of the data contained in the matrix. The final blocks group a number of sentences linked by specific words which express a potential topic.

The usual statistical method (ACP or AFC [5]) give a larg number of plans containing a wide variety of cases which seriously complicate the analysis (problem of graphic representation).

The high frequency words are the most used words in the French language and the low frequency words correspond to some very precise terms in the text. According to this, it is possible to underline three zones [6] (figure 3).

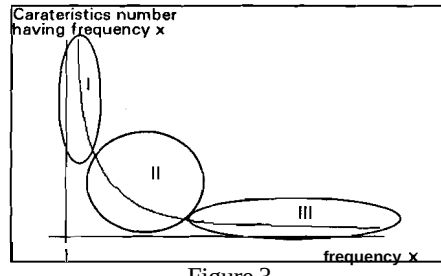

Figure 3

Zone I : commonplace words, generally terms used in the speech.
Zone II : words which are likely to provide interesting information.
Zone III : words which are not likely to provide interesting information.

Moreover these matrices have a particularity : they are « very empty » (few presence of statements). Because of the non gaussian type of this distribution [7], the classic methods of data analysis are applicable but influenced by this phenomenon (ACP, AFC etc..). We thought to developp a new method which allows to define the different zones and to separately analyze each part in order to determine groups of words in the sectors of the curve where the frequencies of these words are relatively not very distant.

Statistical analysis

We consider a set X of N sentences and a set Y of M words, we define the matrix
$A = \{a_{ij}\}$, i=1,2,....,N; j=1,2,...,M as follows:

a_{ij} = 1 if the sentence i possesses the jth word

$a_{ij} = 0$ if not

Given A, the problem consists in maximizing the following function :

$$F(n,X_n,Y_n)=\sum_{k=1}^{n}\sum_{i\in X_k, j\in Y_k} a_{ij}+\sum_{k=1}^{n}\sum_{i\notin X_k, j\notin Y_k}(1-a_{ij})$$

where n is a given number of classes, $X_n=\{X_1,X_2,...,X_n\}$ is a partition of the N sentences into n classes, and $Y_n=[Y_1,Y_2,...,Y_n\}$ is a partition of the M caracteristics into n classes.

In fact we can give a trivial definition by this sentence :

The seriation is maximum when the number of 1 inside the blocks and the number of 0 outside the bloks are maximum.

The theorical maximum of such a function is given by

Max (F) = M x N

Hence, MARCOTORCHINO [3] has defined a block seriation coefficient associated with the GARCIA-PROTH [4] function as follows :

$$S=\frac{F(n,X_n,Y_n)}{MN}$$

This ratio can varied in the interval [0,1]. The value 1 corresponds to a perfect block seriation. in which the blocks contain 1s and the matrix. outside the blocks. contain 0s. Lets us mention here that this objective function is perfectly defined and allows us to obtain a simple and efficient quality of the result. This ratio is recognized by all the block seriation scientific cornmunauty. F. MARCOTORCHINO rewrote the problem as linear. The linear approach is perfectly resolved by the linear programming. The main difficulty remains the computer time used for analysing the matrix. In fact the calculation time of matrix (NxM) is really acceptable for N values less than 30.

The matrix is treated by the software DATABLOC in order to extricate some groups constituted with sentences and words representative of the topics evoked by the firm manager.

Interpretations and results

The interpretation of the results (on a set of 5 analyzed interviews) can be obtained according to two levels, statistical interpretation and interpretation of interviews content (differentiation between content and structure of the speech). From a statistical point of view the number and the size of classes which appeared during the seriation, give an idea of the global structure of the speech.

Some classes rich in words, but still regrouping a limited number of sentences (rectangular classes) are significant of long sentences in the reference text. This could characterize some long sentences using an extended vocabulary and / or dealing with numerous themes ("talkative" firm manager), but this does not enable us to provide a finer interpretation in term of sense.

It is the same as several classes with small dimensions ("square" classes), meaning of conciser sentences, using a more restricted vocabulary and / or dealing with a minimum of topics ("not very talkative" firm manager).

It is difficult to generalize on the interviews analyzed such statistical interpretation because of numerous bias which could be brought to this kind of analysis. One of the most important bias could be the behavior of the interviewer himself, who could influence directly the predisposition of manager to answer and the volume and the nature of his answers.

However some works are carried out to find indicators allowing to quantify the speech structure.

On the other hand a semantic interpretation is possible by the visualization under an hypertext form of classes obtained by the seriation and of sentences from source text to which they do refer.

Indeed, a class is composed of a set of words under their reduced form and of numbers of sentences in which they take place. So this is not immediately exploitable from a speech interpretation point of view.

But a visualization tool allows us to consult the classes and their content in regard of reference sentences in the speech (figure 4).

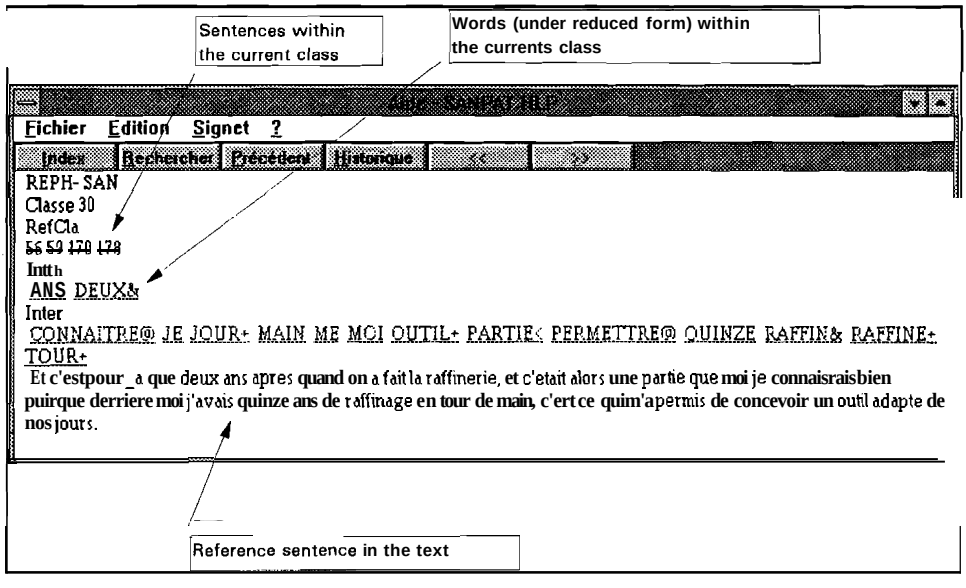

Figure 4

Referring back to the primary document, as an indispensable stage of any reading assistance, enables here to establish the nature of links between the different sentences of a same class.

Some of these links are relative to the polysemia of some words: several sentences in a same class including a same word having a different meaning in each of them, and therefore establishing a "statistical" link between these sentences and not a "semantic" one. This could have been avoided by a process getting rid of any ambiguities, implying the use of computerized synonyms dictionary and a morpho-syntactic linguistic study [8]. But the extent

of the treatment would have disgressed from the setting of this survey which is the feasibility of a relatively simple treatments chain on microcomputer (type PC- 486).

But other links allow to make a connection of ideas shown throughout the speech, and to give the themes or important preoccupations of the text.

Conclusion

In conclusion, it is acceptable to specify what could bring this tool of statistical analysis compared with indexed inconveniences.

It is obvious that what is suggested here constitutes a tool for reading assistance, which does not exempt from reading the primary document [9]. This philosophy is in fact adopted in the visualization under an hypertext form of classes compared with original text.

Moreover, a precise analysis of the references contents is only possible with an expert of the considered domain, who knows how to evaluate the relevance of revealed themes in text. This is in accordance with the process carried out in our laboratory within the framework of the Technological Survey and Competitive Intelligence [10].

The slowness of the global process is to be considered parallel to the accessibility of used equipment (standard IBM PC).

It is to be specified that, in the case of interviews representing the sample of work, the authors are completely conscious of the aleas inherent to a semi-directing interview, and to the numerous factors which could influence the quality of relationships between the interviewer and the interviewed. Therefore it has been possible to isolate, with this tool, some interesting concepts in relation with innovation.

Such an analysis offers a reproducible homogeneous reading from an interview to an other, which is not always possible with a sequential reading of several interviews by a same person.

A good idea of the internal structure of a text is also provided. The structure and the number of classes reflect the physiognomy of the text.

Finally, one important element : it allows to link, in a statistical way, some parts of the text "physically" distant, and which therefore could have been overlooked by a sequential reading. The semantic analysis will next determine the value of these links in the setting of the studied domain.