Semiotics and Modeling Computer Classification of Text with Genetic Algorithm : Analysis and first Results

Vincent Rialle*, Jean-Guy Meunier**, Sofiane Oussedik*, Georges Nault**

* Laboratoire TIMC-IMAG, UMR CNRS 5525, Université Joseph Fourier, Grenoble.

E.mail : Vincent.Rialle@imag.fr.

** Laboratoire d'Analyse Cognitive de l'Information (LANCI). Université du Québec à Montréal.

(Proc. 1997 Int. Conf. Intel.t Systems and Semiotics : A Learning Perspective (ISAS'97). Nat. Instit. for Standards and Technol. Special Public. 918, Gaithersburg, Maryland, p. 325-330.)

Abstract. Computer engineering proposes the construction of complex systems by dynamic prototyping (Buddle and Bacon, 1992). But this prototyping cannot be inductive and purely considered as a trial an error process. To be successful, one must possess an underlying hypothetical model (Marr, 1982) of what are the functions of the system. If these functions relates to physical tasks, such as sensing temperature, manipulatiing objects, etc., the desired behavior can be observed, and a model can be built. Conversely, if the functions of the system are to be applied to semio-informational tasks, such as *language translation, information retrieval, hypertext navigation, text generation,* etc., the interpretative behavior is not readily observable. Now, as any other computer systems, these systems are symbol manipulation machines (Newell ,1980). They must also manipulate input and outputs, but, in themselves, these data are *semiotic* objects, and not physical ones. These systems manipulate objects that have to be *interpreted* by some cognitive agent. In other words, systems that manipulate physical objects require a model of the physical word, while systems that manipulate informational objects require a *semiotic model*. In this paper, we illustrate how a semiotic model can help in the conception, the modeling, and the experimentation of a semiotic behavior such as Computer Assisted Reading and Analysis of Text (CARAT), and how this model has called upon the Genetic Algorithm (GA) theory to realize some of its aspects.

I. Presentation of CARAT

I.1 General presentation

Computer Assisted Reading and Analysis of Text is the computer technology that offers readers an asssistance in attaining some aspects of the informational or semiotic content of a text (discursive, lexical, hypertextual, thematic, stylistic, etc.). So, CARAT definitely relates to interpretative actions. There is in no way a robot that reads or understand a text by itself.

One the classical models of text interpretation is the philological one[1]. Through the centuries, thousands of readers, exegetes, and interpreters have practiced this method. Because of the quality of its principles, it has acquired compelling recognition, and the weight of its experience. The basic principle of philological perspective is that one can construct relatively systematic procedures capable to ensure rigor in text interpretation. As a matter of fact, *philology is an instanciation of an interpretative semiotic process applied to the processing of textual signs*. It takes sets of signs (a text) as its input, then classifies, categorizes them, explores and selects them, and produces a new set of signs - the commentaries - as its output. This interpretation process can be translated functionally in terms of (a) *inscription,* (b) *classification,* (c) *exploration, and* (d) *configuration, of information* (Seffah and Meunier, 1994). In its principles, three important dimensions can be emphasized : text reading and analysis is a *systematic, dynamic and plastic behavior. Systematicity* pertains to the controlled processing of information; *dynamicity* concerns the interaction of the analyst with the text; and *plasticity* allows the constant renewed interpretation of the text.

In order to respect this particular type of interpretation process, a computer model must rely on an open architecture. It must allow an information processing flow that is systematic dynamic, and plastic. Each processing will be built out of interactive advances and restarts which sometimes are autonomous, sometimes are interrelated, but which all aim at assisting the reader and analyst in penetrating the content of the information. Hence, again a CARAT system is not a robot reader, but a faithful assistant in reading and analyzing texts. In this perspective *CARAT is defined as the set of serial or parallel operations which, with the assistance of the computer, construct interpretative paths in which each moment produces a new textual object to be classified, explored and configured*.

I.2 CARAT and classification

---

[1] A certain number of researchers using information technologies are beginning to place themselves in this philological perspective (cf. Thrane *et al.*, 1992).

There exists an infinity of possible CARAT processing flow. Each text, for each person, can be read and analysed in so many ways. One can, for instance, *inquire on a particular theme, paraphrase and summarize a specific segment, study the lexicon, evaluate the style, retrieve information, build a thesaurus or an index of the content, and so forth*. Among all the operations at work in each of these processing, we will study more particularly the classification process. This process is important since interpretation always requires some type of classification of the incoming signs, symbols or information.

In the field of text processing, there exists many strategies of classification. Some classical strategies : a) are of logico-symbolic type (eg (Hobbs, 1993; Sowa, 1991) or semantico-linguistic (Rastier, 1987, Bertrand-Gastaldy *et al.*, 1987, 1993). Others are statistical (Church and Hanks, 1990; Reinhert, 1994; Lebart and Salem, 1994; Pustejovsky, 1991; Wilks, 1996; Salton, 1989; etc.). Albeit very *systematic* these approaches lack *dynamicity* and *plasticity*. Learning is limited, and they are very weak on processing a constantly ever-changing informational input, as it is often the case with textual data (for instance on the World Wide Web). Finally, some can be referred to as "emergent computation" models (Forrest, 1991). They include Markovian fields (Kindermann and Snell, 1980; Bouchaffra and Meunier, 1995), connectionnism (Rumelhart, 1986; Salton and Buckley, 1994), and Genetic Algorithms as shown in this article. Besides their properties of statistical strength and generalization, they are *systematic* like any other clustering strategies, dynamic and plastic (learning is possible).

Our purpose here, is mainly to show how the Genetic Algorithm approach to classification can be applied to the problem of semiotic interpretation of text, and most of all in the context of CARAT technology. Although validation and experimentation of the GA approach is not the main purpose of this paper, which is modeling, some initial experiments will be reported.

II. Genetic Algorithms

II.1 General presentation

The GA approach takes its inspiration from research done on adaptive systems. This research sees such a type of system as an *agent that applies* to a domain (called the environment) specific operations which allow him to act upon it in the most efficient manner . This principle is of course based on the assumption that an adaptive system is able to detect, or extract, from its heteroclite domain any regularities which concern it, and vis-à-vis which it must construct a plan of adaptation.

"The adaptive plan determines just what structures arise in response to the environment, and the set of structures attainable by applying all possible operator sequences marks out the limits of the adaptive plan's domain. " (Holland 1992: 4)

In other words, a strategy of adaptation is the best *plan of action* that a system could put into place in order to identify the structures of its environment. In a more traditional sense, it uses some type of pattern recognition strategy in order to adapt.When applied to the field of genetic reproduction of species, this strategy of adaptation consists in finding, for a given environment, groups of individuals chromosomically best adapted. When constructed in a formal model, this strategy translates into an algorithmic model called *genetic algorithm*. The notion of genetic algorithm, presented for the first timeby John H. Holland in 1975 (Holland, 1975; Holland, 1992), was considerably developed during the 1980's and 1990's (Goldberg, 1989; Rawlins, 1991; Varela and Bourgine, 1992; Michalewicz, 1994).

The main function of this algorithm is the production of a population of individuals, out of an original population, best adapted to an environment which represents the constraints and particularities of the problem dealt with. The degree of adaptation is evaluated by means of a fitness function f. So, the GA is based on:

- an incoding of information, situations, problems and solutions, in the form of strings of building blocks, each string being able to be broken between each block, in the exact image of chromosomes which constitute veritable lists of characteristics of an individual. This incoding usually takes the form of a highly structured binary string, of a fixed or variable length according to the type of problem;

- the capacity to reproduce such strings in large number, which metaphorically relates to the sexual reproduction;

- the existence of a faculty of adaptation (simulated by the function f) which permits the evaluation of the quality of each individual created by the algorithm.

II.2 Basic cycle of a GA

In practice, a population P^0 of potential solutions (the chromosomes) to the problem to be treated is generated at the initialization step. Then the following standard cycle, also called *genetic search*, is reapplied :

INITIALIZATION

If (stop test not verified) then

begin

EVALUATION ; SELECTION ; REPRODUCTION ;

REPLACEMENT

end

At any given moment t, the population is: $P^t = \{ a_1^t, \ldots, a_p^t \}$, where a_i^t stands for the candidate solution a_i at cycle number t. The main steps can be briefly described as follows :

1) EVALUATION: The elements of P^t are rank ordered from the most to the least fitted according to the selection probability $Prob_S$:

$$Prob_S(a_i^t) = f(a_i^t) / \left(\sum_{j=1}^{p} f(a_j^t)\right)$$

2) SELECTION: The elements which best satisfy the constraints or characteristics of the solution sought are selected according to the selection probability, and arranged by couples in order to prepare the next step.

3) REPRODUCTION: *genetic operators* are then applied to this population of élites, called *parents*, in order to obtain an intermediary population P'^t. These operators permits the creation of new strings, among which some should have better fitness properties than their parents. Two genetic operators are generally employed: *crossing-over*, which allows the production of two new elements from two parent elements (Fig. 1), and *mutation*, which allows the creation of new solutions that would have been impossible to obtain by simple crossing. Mutation consists of a random selection of one of the bits of the chromosome, and to change its value with a pre-defined probability (probability of mutation) (Fig. 2).

4) REPLACEMENT: This is the generation of a new population by replacing the worst elements of the previous population P^t by the best of P'^t. This new population possesses at least as many and sometimes more of the characteristics of the solution than the preceding generation.

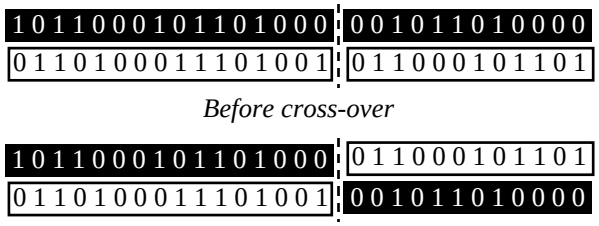

Fig. 1. Diagram of crossing-over

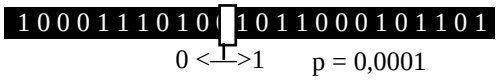

Fig 2. Mutation

This cycle is iterated a lot of times until a generation of optimal solutions is obtained, and from which only the bests will be retained. In certain contexts it is the entire population which stands in lieu of the solution.

III. Genetic Algorithms and CARAT

III.1 The object of the model

As previouly said, CARAT can fan out in many different processing flows. In this section, we propose a modeling, by means of a genetic algorithm, of the classification moment of two specific processing flows : the first flow aims at giving the reader hints on the semantics contexts of particular words ; the second flow aims at automatically suggesting hypertext links among segments of texts. So, the main concepts that define GAs will be translated in the terms of classifying segments of texts, and this theoretic exposé will be followed by a presentation of some experimental results.

In the CARAT context, genetic algorithms consists of finding, amongst different segments of a text, which one offer some regular structures or form classes of regularities. The GA is seen as a process of classificatory treatment which identifies segments of text containing some identical "type" of information. These segments are most often *pages*, and contains *unifs* (for units of information), which are simple or compound words, lexemes, etc. The determination of unifs and segments can be done by means of specialized computer text analyzing programs such as SATO, BOOKMANAGER, SPIRIT, OPEN TEXTE, NATUREL, etc. Here one creates a lexicon, a *linguistic markup*, a *tagging*, etc.

So, the original text is transformed into an set of segments containing only a balanced and controlled choice of units of information. A procedure identifies the presence or absence of each unif in each segment, and builds the following matrix (Tab. 1) of *n* segments by *m* unifs.

	unif 1	*unif 2*	. . .	*unif m*
seg. 1	Pre(1,1)	Pre(1,2)	. . .	Pre(1,m)
seg. 2	Pre(2,1)	Pre(2,2)	. . .	Pre(2,m)
. . .	. . .	. . .	. . .	. . .
seg. n	Pre(n,1)	Pre(n,2)	. . .	Pre(n,m)

Table 1. Matrix segments-unifs

One particular segment is represented by a line vector of binary numbers given by the predicate Pre(i, s) : 0 for absence, 1 for presence.

III.2 Genetic classification

The object of the model is to assign each segment to a specific and unique class. The assignment of a segment to a class is called *classing*, and uses a *classifier*, whereas the process of research of the best classing (i.e., the best classifier) should be called *induction of classification.*

III.3 Set definition and initial population

In order to introduce, in the context of CARAT, the concept of *population of individuals* (Fig. 3), and particularly the one of *initial population*, we must appeal to three important sets :

a) the set **T** corresponds to the set of segments of the text. Let $\mathbf{T} = \{S_1, \ldots, S_n\}$

b) the set **K** of classes. Let $\mathbf{K} = \{C_1, \ldots, C_{NbC}\}$

Where NbC represents the number of classes. A class is a set of segments which are not too distant one from another according to the function of adaptation defined below. At the initialisation setp, the number of classes is arbitrarily chosen large (it is equal to the number of segments n) in such a way as to give the process the freedom to construct as many classes of segments as possible. It is the purpose of the function of adaptation to reduce, during the genetic search, this number to an optimal value. The number of cycles is also arbitrarily fixed at a fairly large value, in the order of one thousand cycles. Finally, the interpretation of each class is devolved upon the user.

c) The set (or *population*) **P** of the individuals (P^t represents the state of the population at time t. It contains a fixed number p of individuals). The elements of **P** are called classifier-vectors, and represents the candidate *classifiers*, *i.e.*, the potential solutions to the

problem of the best classing. In other words, an individual *encodes* a *tentative solution* for classing segments. The size of the classifier-vectors is n ; the position number i corresponds to the i-th segment in the text, and contains an integer equal to the number of the class to which the segment belongs.

At the outset, the classifier-vectors are randomly built to produce the initial population P^0.

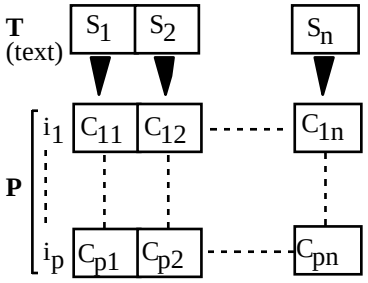

Fig. 3- relations of sets T, K and P.

The genetic search has the task of carrying out a considerable number of modifications to the classifier-vectors, such as recombinations and mutations, in an attempt to find the best one according to the function of adaptation.

The j-th individual of P corresponds to the vector :

$V_j = (C_{j1}, \ldots, C_{jn})$, $C_{ji} \in \mathbf{K}$

An individual or chromosome could, as well, be seen as a set of couples which associates each segment to a unique class. The individual represents also a function which plunges the set **T** of segments into the set **K** of classes.

Example :

Let $\mathbf{T} = \{S_1, S_2, S_3, S_4\}$; $\mathbf{K} = \{C_1, C_2, C_3\}$

Example of classifier-vector : $\mathbf{V} = (2, 1, 3, 2)$

Interpretation: the first segment belongs to the class C_2, the second to the class C_1, the third to the class C_3, and the last one to the class C_2.

III.4 The function of adaptation

III.4.1 Finality

The function of adaptation must evaluate the intrinsic value of an individual, and hence the quality of the classing that it codes. This function is defined over the set of individuals and gives a real value.

In an ideal classifier *a*, segments are grouped into compact classes. This quality is characterized by the fact that displacing a segment i from one class to another (*i.e.*, changing the class number located at position i in the classifier-vector) could result only in a decreased value of $f(a)$. The individuals selected for reproduction are those which posses the best values by f. Therefore, this function directs the entire process of the genetic search, and the quality of the whole GA model depends essentielly on it.

III.4.2 Choice of a criterium of similarity

It behoves the GA designer to conceive the most efficient function of adaptation . In particular, this function must be most discriminating. This discriminating feature involves two complimentary aspects :

- the evaluation of internal cohesion of classes, *i.e.*, the degree of similarity of segments within each class;

- the evaluation of the differentiation between classes of the classifier-vector, *i.e.*, the degree of contrast existing between classes.

The function of adaptation we propose is based upon the *score of Jaccard* as criterium of similarity. This score uses only the property of presence or absence of unifs in segments, and constitutes a common measure for evaluating the similarity of textual documents in the case of information research. It has been used notably for indexation (Gordon, 1988). However, there do exist a few other criteria (Salton, 1989).

The Jaccard score of a couple of segments (X_j, X_k), notated $\mathrm{Sim}(X_j, X_k)$, is equal to the proportion of unifs common to both segments (notation : $|X_j \cap X_k|$) relatively to the total number of unifs present in the two segments: $\mathrm{Sim}(X_j, X_k) = \dfrac{|Xj \cap X_k|}{|X_j \cup X_k|}$

where : |A| represents the cardinal of the set A.

- the *internal cohesion* of a class is evaluated by a *coefficient of internal cohesion* noted as $IC(C_i)$. The coefficient is the balanced sum of the similarity of segments taken two by two in this class. It is defined by:

$$IC(C_i) = \frac{1}{N(i)} \times \sum_{\substack{Xj, X_k \in C_i \\ j \neq k}} \mathrm{Sim}(X_j, X_k)$$

$$CI(C_i) = \frac{1}{N(i)} \times \sum_{\substack{Xj, X_k \in C_i \\ j \neq k}} \frac{|Xj \cap X_k|}{|X_j \cup X_k|}$$

N(i) is the number of combinations of the segments of the class C_i taken two by two.

- the *differentiation* of the classsing is evaluated by the *coefficient of external dissimilarity*, noted as $ED(C_i)$, and computed for a class C_i in relation to all other classes. It is defined as follows:

$$ED(C_i) = 1 - \frac{1}{NC(i)} \times \sum_{Xj \in C_i, X_k \in CC_i} \mathrm{Sim}(X_j, X_k)$$

$$= 1 - \frac{1}{NC(i)} \times \sum_{Xj \in C_i, X_k \in CC_i} \frac{|Xj \cap X_k|}{|X_j \cup X_k|}$$

CC_i is the complimentary set of C_i. This set contains all of the segments that do not belong to C_i. NC(i) is the number of couples (X_j, X_k), X_j belonging to C_i, and X_k belonging to CC_i.

Finally, the function of adaptation f is equal to the sum of these two coefficients.

$$f(\text{individual}) = \sum_{i \in [1, NbC]} (IC(C_i) + EC(C_i))$$

All the inter-segment similarities are computed at the initialisation step, before the GA search.

To summerise. The genetic search runs a number of cycles equal to the maximum number of cycles determined at the start-up of the GA. When the cyclic processing has finished, the resulting population represent the best classifiers that the GA could produce. The last step is to select from this population the classifier-vector which gives the greatest value for f. At the end of the algorithm, a certain number of classes are empty. This is both expected and hoped since the number of classes was arbitrarily fixed at a high number. Thanks to the function of adaptation, the algorithm converges towards an optimum number of non-empty classes.

III.5 Experimentation results

The experiment was carried out on a textual sample drawn from *Spirale*, a Belgian review on Education Sciences. The GA was developed using Matlab, and was integrated into a software platform, Aladin (Seffah and Meunier, 1995), developed at LANCI for the CARAT approach.

The probability of crossing-over was fixed to 0.8, and the one of mutation to 0.05. The number of individuals in the initial population was 100, and the number of generations (or iterations) was 300.

The text was partitioned uniformly into 54 segments of 50 words each, the end of a segment being determined as follows: fifty words are counted from the beginning of the segment and then any words remaining up to the next point are added to the initial fifty words to constitute the whole segment. We had at our disposition a lexicon composed of 1701 words of which the number was restricted to 1360 roots after a preliminary process of lemmatisation. So, the size of the text matrix was (1360 x 54).

The genetic search decreased the number of classes from 54 to 24. Here is a short sample of interpretable results. For instance, Class 4 contains the following segments 8 and 21, in whicn underlined italic words are common unifs determined par the GA :

Segment 8 : « At last, Joëlle Delatte is grappling with the problem of books for blind and sight-impaired *children*, which would seem to be the preoccupation of at least some editors who have recently proposed specially designed albums for them. This production is characterized by a certain diversity if however unified by the prudence of their approach : *Children's literature* does not represent all the *reading* nor all the *literature*, but it does exist with a sufficiently rich past and present. »[2].

Segment 21 : « To the teacher convinced of the importance of *reading literature* there exists a question of choice of texts and how to transmit them. There exists at this level no ministerial propositions nor lists of lists of books as there are for colleges; nor a specialized university teaching tradition to define the methods; the initial and ongoing training is incongruous and left largely to one's_own initiative. In effect, the teacher who chooses his own texts and his own course of action, wittingly or not, is putting into action his personal conception of the culture and the role of the school in the education of the *child*. »[3].

Within these two segments, three unifs have been included in the same class : *child, literature, reading*. This result might facilitate the work of a user facing the problem of reading and analysing a large text, by suggesting a precise relation between the segments. A more in-depth analysis of the results would of course require the help of a terminologist or a specialist of the field. Such help is just as indispensable at the moment of preparation of the text matrix as it is at the end for the analysis of the results.

So, the genetic algorithm has built classes that classifies segments of text that offer some lexical similarity. It has done this in a systematic, dynamic and plastic manner. From these classes of segments the processing flow can then whether choose a particular word and see the class of similar segments in which it operates (its particular semantic contexts) or choose to built a hyperlink between two similar segments.

Conclusion

The use of the genetic algorithm that we have applied to the analysis of a text is still at the experimental observation stage. But it is, however a very promising territory and very flexible, which combines coding, the processing of data, computation of probabilities, artificial intelligence, and the genetic mode of evolution. The variants of the model presented are numerous and are linked to the diversity and richness of the genetic operators, to the multiple ways of coding the solution, and to the conception of the function of adaptation.

Research perspectives are situated around the most in-depth study of the genetic algorithm applied to CARAT and to the comparison of results obtained with

---

[2] « Enfin, Joëlle Delattre aborde le problème des livres pour enfants aveugles et malvoyants, dont certains éditeurs semblent d'ailleurs se préoccuper en proposant maintenant des albums spécialement conçus pour eux. comme on le voit, une certaine diversité caractérise cette livraison, mais son unité nous semble résider dans la prudence des approches : La littérature de jeunesse ne représente ni toute la lecture ni toute la littérature, mais elle existe, avec un présent et un passé suffisamment riches. »

[3] « Se pose alors, à l'enseignant convaincu de l'importance de la lecture littéraire, la question du choix des textes et des modalités de transmission. En effet, il n'y a à ce niveau ni propositions ministérielles de listes d'ouvrages, comme le collège ; ni tradition d'enseignement universitaire spécialisé pour définir de méthodes ; la formation initiale et continue est disparate, largement laissée à l'initiative de chacun. En fait, l'instituteur, qui choisit ses textes et ses démarches, sciemment ou non, met en jeu toute sa conception personnelle de la culture et du rôle de l'école dans la formation de l'enfant. »

other classifying systems that are presently being developed, such as *simulated anealing* and the ART neural network.

We believe that the modeling of CARAT by genetic algorithms allows us to foresee solutions to a certain number of problems of processing textual information that require classification tasks. On the one hand, classification allows regrouping the segments of a text in terms of an optimization of the similarity level for relating segments of information. On the other hand, by the nature of its mathematical structure of the topological type, the GA permits the processing of a body of text that is in constant evolution.

This algorithmic strategy is applicable to diverse types of textual information processing systems, such as terminological classification, thematic extraction in text. Automatic generation of hypertextuel relations.

Thanks

The present work is part of a Franco-Québecois project "Les classifieurs émergentistes et le traitement de l'information" under the tutelage of, for France, le Minisère de l'Enseignement Supérieur et de la Recherche (MENESRIP, Délégation à L'Information Scientifique et Technique, programme "Ingénierie linguistique et de la connaissance"), and le Ministère des Affaires Etrangères; for Quebec: Gouvernement du Québec - Ministère des relations Internationales. We extend a special thanks to these trustees for work and cooperation that they have permitted.

We would also like to thank the entire team at LANCI and in particular M. Nyongwa and I. Biskri for their advice and technical support.

References

Bertrand-Gastaldy, Paquin, L. C., Meunier, J. G., (1993). L'analyse de texte par opposition à la gestion des documents, ICO, vol. 4, 12-18

Bertrand-Gastaldy, S., Meunier, J. G., and Lebel, H. (1987). A Call for Enhanced representation of Content as a Means of Improving Online Full-Text Retrieval. *International Classification* 14 (1), 2-10.

Bouchaffra, D. Meunier, J.G. (1995). A Markovian Random Field Approach to Information Retrieval, ICDAR, *Third Int. Conf. Doc. Anal., and Recog*. IEEE, Computer Society Press. Vol 2, 997-1003.

Buddle, R., and Bacon, P. (1992). *Prototyping: an approach toevolutionary system development*. Springer Verlag

Church, K. W., and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics 16* (1), 22-29.

Davis L. (1991). *Handbook of Genetic Algorithms*, Van Nostrand Reinhold. New York.

Delany, P., and G. Landow (ed.) (1993), *The Digital Word: Text Based Computing in the Humanities.* Cambridge, Mass: MIT Press.

Delany, P., and Landow, G. (1991). *Hypermedia, and Literary Studies.* Cambridge: MIT Press.

Dupoirier, D. (1994). *Technologie de la GED*, Paris, Hermès.

Feldman, J. J. (1988). Connectionnist Representation of Concepts. In J. A. Waltz David & Feldman (ed.), *Connectionist models, and theirimplications: Readings from Cognitive Science* Norwood: AblexPublishing.

Forrest D. (ed.) (1991). *Emergent Computation*, MIT Press, Cambridge, Massachusetts.

Goldberg D. E. (1989). *Genetic algorithms in search, optimisation, and machine learning*, Addison-Wesley. Reading, Massachusetts.

Gordon (1988), Probabilistic, and genetic algorithms for document retrieval. Communications of the ACM, 31(10): 1208-1218.

Hobbs, J. (1993). Intention, Information, and struture in discourse. In *Proceedings of the Nato Advanced Research Workshop on Burning issues in Discourse*, Maraka, Italy, 41-66.

Hoffmeister F., and Bäck T. (1992). Genetic Self-Learning, in Varela F. J., and Bourgine P. (ed.). *Toward a Practice of Autonomous Systems*, MIT Press. Cambridge, Massachusetts, 227-235.

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

Holland J. (1992). Genetic Algorithms, *Scientific American*, July, 66-72.

Kindermann R., and L. Snell (1980), Markov Random Fields, and their applications, *Contemporary mathematics*, AMS, Vol. 1.

Kintsch, W. (1988). The role of knowledge in discourse compr_hension: A construction-integration model. *Pryshological Review 95* (163-182),

Lebart L., Salem A. (1994). *Statistique textuelle*, Dunod, Paris.

Marr, D. (1982). Vision: A Compuational Investigation into Human Representation, and Processing of Information, San Francisco, Freeman Publ.

Michalewicz Z. (1994). *Genetic Algorithms + Data Structures = Evolution Programs*. Springer-Verlag, Berlin Heidelberg.

Newell, A. (1980), Physical Symbol systems.*Cognitive science*. 1984: 2, 135-183.

Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics*. 17(4).

Rastier, F. (1987). Sémantique et Intelligence Artificielle, *Langage*, vol. 87.

Rawlins G. J. E. (ed.) (1991). *Foundations of Genetic Algorithms*, Morgan Kaufmann, San Mateo, CA.

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

Rumelhart, D. E., and Mc CLelland, J. (1987). *Parallel Distributed Processing :Explorations in the Micro Sctucture of Cognition, 2 vol.* Cambridge: MIT Press.

Sabah, G. (1989). L'I. A et le langage vol 2. Paris: Hermès.

Salton G., and McGill, M. (1983). *Introduction to models of Information Retrieval*, New York: Mc Graw Hill.

Salton, G. (1989). *Automatic Text Processing: The transformation, Analysis, and Retrieval of*

Information by Computer. . Reading, M. A: Addison Wesley.

Salton, J., Buckley A. C. (1994). Automatic structuring, and retrieval of large text file*Com. of the ACM, 37, (2), 97-107*.

Seffah, A., and Meunier, J.G. (1995). ALADIN: Un Atelier génie logiciel orienté objets pour l 'analyse cognitive de textes. In Bolasco. S, Lebart, L. Salem., A. : *Analisti Statistica dei Dati Testuali*. JADT, 1995, Rome, CISU, VOL II, p. 105 -113.

Sowa, J. F. (1991). *Principles of semantic networks*, San Mateo: Morgan Kaufman.

Spark-Jones, and Kay. M. (1973), *Linguistics, and Information*, Science Academic Press Londres.

Thrane T., Olsen J., Jansen S., and Prebensen H. (1992). *Discourse Analysis*. Copenhague : Museum Tuscalanum Press.

Varela F. J., and Bourgine P. (ed.) (1992). *Toward a Practice of Autonomous Systems*, Proc. 1st Europ Conf Artificial Life, MIT Press, Cambridge, MA.

Virbel, J. (1993). Reading, and Managing Texts on the Bibliothèque de France Station. In P. Delany, and G. Landow (ed.), *The Digital Word: Text Based Computing in the Humanities.* Cambridge, Mass:MIT Press.

Wilks, Y. A. (1996). *Dictionaries, Computers, and Meaning.* Cambridge: MIT Press.

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