User Navigation in the IRS Conceptual Structure through a Semantic Association Function

MARISTELLA AGOSTI[1] AND PIER GIORGIO MARCHETTI[2]

[1] *Dipartimento di Elettronica e Informatica, Università di Padova, Via Gradenigo 6/a, 35131 Padova, Italy*

[2] *European Space Agency, Information Retrieval Service (ESA/IRS), Via Galileo Galilei, 00044 Frascati, Italy*

This paper addresses the methodological aspects of designing a semantic association function in a hypertext information retrieval environment and model. Through this function a part of an associative information retrieval model is designed and implemented. Initially the paper provides reference to previous efforts in associative information retrieval and automatic search formulation and re-formulation operations. The semantic association function is a function which belongs to a new functional hypertext information retrieval model that has been previously justified and presented. For this reason, the meaning and aim of the semantic association function are also recalled and reported. The central part of the paper presents motivations and justification for the design and implementation of the semantic association function in an experimental hypertext information retrieval interface as opposed to a traditional online information retrieval system.

Received December 1991

1. INTRODUCTION

In the past few years, the availability of first-generation hypertext systems has been considered with great attention by the information retrieval community. New information retrieval applications integrating hypertext have been proposed and studied (see for example Refs 7, 8, 10, 14, 15, 18). Hypertext allows easy linking, browsing and navigation operations on pieces of information of different nature like chunks of text, images and sound or video. The hypertext capability considered most important for a new generation of information retrieval systems would be the possibility to also link documents and concepts. We make this statement on the assumption that the most elementary way of thinking and learning is by association.[9]

This hypertext capability can therefore elicit document-to-document, concept-to-concept, or concept-to-document associations. However, this capability which appears so attractive in supporting a concept-centred user–system interaction, still suffers from the well-known hindrance of hypertext systems brought about by disorientation and cognitive load[2,23] for the end-user.

With the assumption that the hypertext approach can lead to a new model for information retrieval,[1] a study produced a new functional model capable of overcoming the major limitations of hypertext systems in relation to information retrieval operations.[4] The new, concept-centred, hypertext information retrieval model incorporates some important information retrieval functions and assists the final user by means of a new type of associative information retrieval; two most important features of such a model are a semantic association and an associative reading function.

The purpose of these two functions is to reduce the cognitive load on the user, giving him the possibility to use natural language words which are automatically related to the concepts the system is able to recognise and to the documents the system relates to those concepts.

In particular, the purpose of the semantic association function is to make transparent and to communicate the meaning the system assigns to the concepts with which the user expresses his information needs. Associative reading, instead, reduces the disorientation of the user by providing a guide for browsing the structure of concepts and the hypertext of documents.

This paper focuses on the methodological aspects of designing a semantic association function in a hypertext information retrieval environment and model. Through this function, a part of an associative information retrieval model is thus designed and implemented. Section 2 provides reference to previous efforts related to associative information retrieval, paying specific attention to early work on term-based retrieval,[12,17] and on automatic search formulation and re-formulation.[16,22,34] The previous work is reviewed in the light of an approach centred on concepts rather than on individual terms, also considering its implications on user–system interaction as defined in Ref. 20. In Section 3, the significance and purpose of semantic association is presented at a functional level as proposed in Ref. 4. Section 4 presents motivations and justification for the design and implementation of the function in HYPERLINE,[4] an experimental hypertext information retrieval interface,[21] as opposed to a traditional online information retrieval system.

2. REFERENCE TO PREVIOUS EFFORTS

Associative information retrieval has been presented in the past as a possible alternative to the issue raised by exact-match retrieval. In exact-match (boolean) retrieval, the information retrieval system produces results which reflect the exact correspondence existing between words and terms in the query and words and terms in the bibliographic references or documents.

It is clear that the way concepts are represented in documents, or in the document surrogates stored in the information retrieval systems depends upon the authors' personal attitudes and on the historical context. Furthermore the technical jargon itself always tends to change. This results in possible unsatisfactory outputs from exact-matching searches. In fact, the retrieved set

may not contain all the possible relevant information and the user may find trouble in re-formulating his query. This problem has been addressed in different ways in the past. The work on associative retrieval refers back to work carried out in the 1960s[17] concerning term associations. Two of the restrictions with which this kind of term association was faced were so strong that the final results often turned out to be quite unsatisfactory. The first restriction was brought about by considering term associations fixed in the database and therefore somehow independent of the context and of the actual user information needs and search goals.

The second restriction arose from the idea of using single terms (quite often single words), whilst for information retrieval users it is more important to recognise concepts, and concepts are better described by multiple words (i.e. terms in a more general sense). In several instances the different results available after term associations in free text and term associations with controlled vocabulary were not immediately evident.[12] Furthermore, certain negative reports on effectiveness of controlled vocabulary usage were based on early works related to information retrieval systems which did not provide for user interaction or in which retrieval was performed via controlled vocabulary only.[13] A lot of work was in fact based on another pioneering paper[19] where a word's significance was related to the word's position in a text. The consequence is a lack of semantic analysis which might in some cases produce unsatisfactory results.

Another view took into account the automatic ranking of documents and query re-formulation, both playing a central role in the design of those information retrieval systems capable of satisfying the user's need for 'relevant' information notwithstanding the inability of the initial query formulated by the user to take into account all the possible term variants, the author's jargon and perhaps even the user's inability in expressing his information needs in a precise manner.[24,32]

The mathematical background was provided by a pioneering work on automatic abstracting and indexing[13] where a word's frequency in a text had been linked to the word's 'significance'. The mathematical formulation evolved in time, incorporating a probabilistic approach towards the identification of relevant documents, assigning 'weights' to terms in the query. In order to keep the model close to reality, a partial dependence between terms was introduced.[31]

Whilst the probabilistic approach does not require any *a priori* classification of the text to be retrieved, the approach that is proposed in the next section requires a preliminary identification of the concepts (indexing) in the text, to be done either manually by experts or automatically.

In large online bibliographic database this is done associating to each document or bibliographic reference a few terms, descriptive of the document itself, from a controlled vocabulary or from a thesaurus.

In this paper it is assumed that the concepts descriptive of the text are extracted either by experts during the manual indexing process or via automatic knowledge extraction algorithms.

3. THE SEMANTIC ASSOCIATION

The semantic association function is part of the functional capabilities of a hypertext information retrieval environment designed to interact with the user who seeks information. The prototype of this information retrieval environment has been introduced in Refs 4 and 5 and is based on the two-level architecture proposed in Ref. 3. The main features of the two-level architecture are:

1. The first level: the collection of documents of interest – the set D.
2. The second level: the conceptual plane on which the semantically related concepts are placed; this is the plane of abstraction where indexing terms used by an IR system can be placed, each term identifies a specific concept.

At the second level there is the set T which represents the universe of possible usable terms. The set S (system terms), a subset of T, is the set of terms used and managed by the prototype. S is the set resulting from the union of all terms used. Two different kinds of terms are used concurrently: the set E and the set C, where:

(i) E (extracted terms) is the set of terms produced by the application of an automatic parsing algorithm to the textual parts of the managed documents; the terms of this set are all the terms extracted by the algorithm that are not included in a list of stop-words or non significant words.

(ii) C (concepts) is the set of indexing terms belonging to an auxiliary data structure used by the prototype. The expression 'auxiliary data structure' describes the structure that supports the representation of the concepts and their relationships used for the representation of the domain of interest of the collection D of documents.

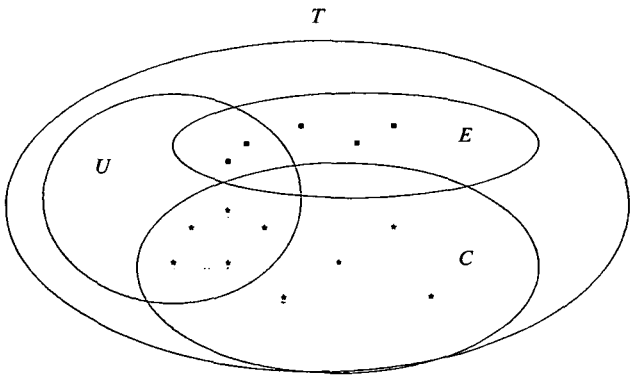
For the first prototype of the hypertext environment, a thesaurus has been used as an example of a complex auxiliary data structure. A thesaurus consists of a semantic structure of indexing terms associated with the documents by experienced indexers. In this context the thesaurus is seen as a repository of human knowledge and ability in concept classification. The fundamental types of semantic relationships expressed in a thesaurus are: scope, equivalence and hierarchical and associative relationships, see Ref. 6 for further details on thesaurus structures; and furthermore:

$$S = E \cup C.$$

The set of all possible usable terms T could contain other auxiliary data structures, but it always contains another set of terms: the set of user terms (U). The elements of the set U are the free terms that the user of the system can insert into his query, that is, the set of terms that are not necessarily present in the set of terms extracted from the documents of the collection or in the set C. It is important to note that the E, C and U sets are not necessarily disjointed, and the following relation can be specified with the set T of all possible usable terms:

$$T \supseteq (E \cup C \cup U).$$

See Fig. 1 for a graphical representation of the different sets of terms in T.



$S = E \cup C$

$T \supseteq E \cup C \cup U$

Figure 1. Different sets of terms in *T*.

The two-level architecture can be used as a frame of reference by the user in the process of query formulation, during documents browsing or concepts navigation. Through the architecture the structure of the auxiliary data is made available to the user, so that he can perceive and navigate through the semantic structure of the indexing terms describing the informative content of the documents.

The functions made available in this environment have been designed taking into account some hypertext functionalities that have been proved useful and relevant for the user, see for example Refs 23 and 29. Furthermore, as a first step towards the integration of hypertext and information retrieval environments, the possibility of using the result of the concept navigation process to construct a search strategy for the user has been provided. This environment provides seven different functions: (1) semantic association, (2) navigation (3) sequential reading, (4) associative reading from a single document or from many documents, (5) backtracking, (6) history, and (7) support for search strategy development. The description of the prototype is as reported.[5,21] This paper concentrates on the peculiarities of the semantic association and in Section 4 gives a formal justification for its design and implementation; in the rest of this section, the purpose and mode of operation of the semantic association are introduced as reported.[5]

The purpose of the semantic association function is to make more transparent and to communicate to the user the meaning the system gives to the terms used in the expression of user's information needs. In the formulation of his query the user can employ natural language, each word given by the user is mapped by the system into a set C_u of semantically similar concepts which is a part of the auxiliary data structure managed by the system. Moreover the set C_u which is provided by the semantic association function operates for the user as an entry point to the auxiliary data structure; the user can via this entry point start interacting with the system. The aim of this interaction is the acquisition of information via concept navigation or document browsing.

The semantic association function operates in the following manner: when the user expresses interest in a topic by using a specific term, a list of conceptually related indexing terms which are concepts of the auxiliary data structure are thus presented.

If the term u that the user initially enters is itself an indexing term, then: $u = c$. Thereby the connections that the term has with the other indexing terms (e.g. the thesaurus relations) are presented to the user who can make use of them to explore the semantic structure of the auxiliary data structure. The terms that are directly connected to the term $u = c$ in the auxiliary data structure C are thus displayed (see Fig. 2). If the term u that the user initially enters is not a concept of C, but element of E, then the set C_u of associated concepts (indexing terms) is constructed by making use of the documents D_u, u would retrieve by means of an inference mechanism on the concepts related to each document during the indexing process.

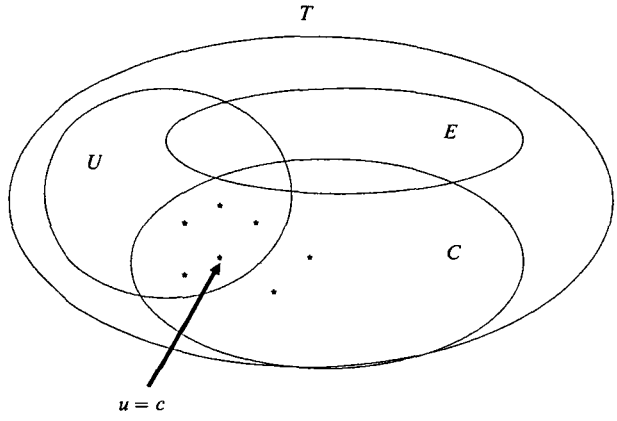


Figure 2. The term *u* is an indexing term.

The procedure is based on the fact that each element of the set E of parsed terms has been extracted during the indexing procedure (parsing) of a set of documents and its relationship to them is maintained. These functional relationships are acknowledged and made usable in an active manner. Thus, a term e is related to a set of documents D_u. Once the set D_u has been constructed by the interface, the latter can construct and make use of the set of indexing terms with which the documents of the set D_u have been associated during the indexing procedure. Since the resulting set of terms of the set C could have too large cardinality, an inference mechanism is applied to reduce the cardinality of the set and to present the user only the set C_u of the most pertinent indexing terms. The way this function operates is shown in Fig. 3, a justification for a possible inference mechanism is given in Section 4.

If the term u that the user initially enters is not a term of the set C nor of the set E, a stemming algorithm proposes an alphabetic list of terms which are morphologically similar to the term u. After the user's choice of a term from those suggested, the interaction continues in one of the two previously presented ways. In the present implementation a list of conceptually related indexing terms is suggested only if the used term u is also an element of the set E. When the semantic association function has completed its operations, the set C_u is available to the user as input to one of the other functions.

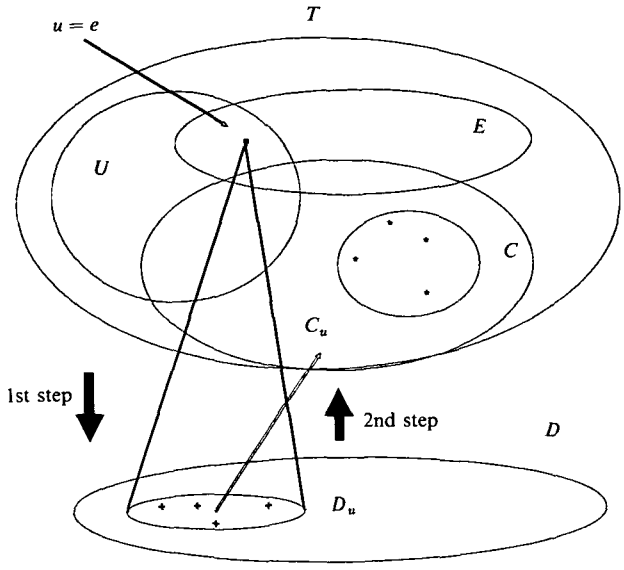


Figure 3. An operational representation of the semantic association function.

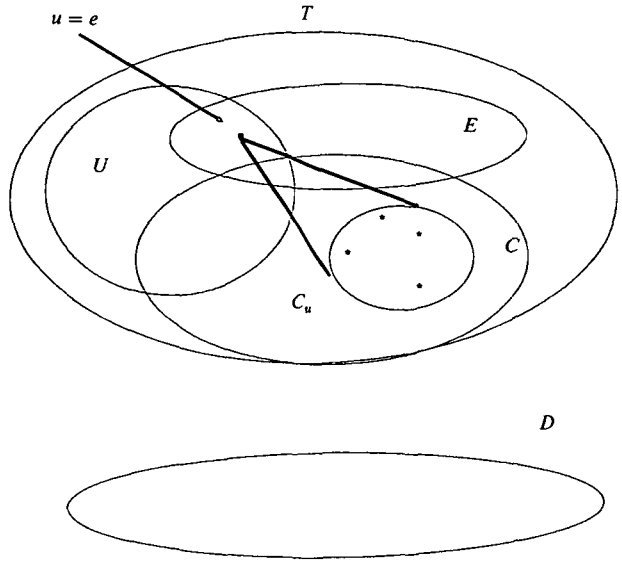


Figure 4. The semantic association function from the user's point of view.

4. A FORMAL APPROACH FOR THE DESIGN OF THE SEMANTIC ASSOCIATION FUNCTION

The semantic association as seen from the user's point of view is a one-step function: the system reacts to input by displaying a list of concepts related to that introduced by the user. For this reason, identification of the sets of documents and concepts the user sees is presented in Fig. 4 from the point of view of the user and not of the system as was done in Section 3. Fig. 4 shows the way in which the semantic association function operates from the user's point of view. In Fig. 4, the symbols introduced hereunder are used to identify the sets of concepts displayed to the user by the semantic association function.

From a system's point of view it is important to remember that in a boolean information retrieval setting the system cannot rely on any information or direct relationships to the concept expressed by the user. The only information the system can use are the relationships to the documents connected to the user's entered concept. Subsequently, the system can infer from the documents that are related to the initial concept a set of concepts and display this set to the user. Due to this feature, the semantic association becomes a two-step function:

1. In the first step the system identifies the set of documents D_u that contain the user-expressed concept u. This being virtually the same as that of a classical retrieval function in a boolean retrieval setting.
2. In the second step of the semantic association the system has to identify the concept or the set of concepts C_u that is associated to the documents in D_u.

Implementation of this second step implies relying on a well-founded inference mechanism. In the following, an approach based on preference relations[28] is presented as being a suitable basis for the inference mechanism.

When the semantic association is performed, the user can potentially examine all the concepts in C_u (see Fig. 4) or, if the size $|C_u|$ of this set is too large, the user might be interested in seeing only a few of them; i.e. the most pertinent to his information needs. Denoting by M the maximum number of concepts that the user wants to examine and by N the size of C_u ($|C_u| = N$), if $N > M$ the question arises as to the choice of the M concepts in C_u in a suitable manner.

The user enters the concept u that represents his information needs. The concepts that are related to u could be chosen in a random manner if all the concepts of C_u were equally informative, or explanatory in relation to the input concept u in the same way, or able to elicit concept association to the same extent. If not, as happens in normal practice, we could introduce an 'order' in C_u using the preference relations approach.[28]

We will assume that as answer to the user input u a set D_u of documents is identified as well as a set C_u of concepts, in the following, to simplify the notation we will make reference only to D and C for the sets respectively of documents and concepts.

Let $d_1, d_2, \ldots, d_v$ be the documents belonging to the set D. The concept c can in general be represented by a binary vector $X = (x_1, x_2, \ldots, x_v)$, where

$$x_t = \begin{cases} 1, & \text{if } d_t \text{ is described by concept } c \\ & (i = 1, 2, \ldots, v) \\ 0, & \text{if } d_t \text{ is } \textit{not} \text{ described by } c. \end{cases}$$

For a description of this kind of IR issue see Refs 26 and 27.

Let X_k be the binary vector that represents the concept c_k, consider the event I_k defined as: 'concept c_k is informative to the user', $k = 1, 2, \ldots, N$. If no specific information is available on the set C of semantically associated concepts, it is natural to assume that the events I_k are exchangeable,[11] thus

$$P(I_k) = p \quad \text{for each} \quad k = 1, 2, \ldots, N, \tag{1}$$

where p is an 'initial' evaluation.

Denoting by $P(I_k \mid X_k)$ the conditional probability that the concept c_k represented by the vector X_k be informative, from Bayes theorem follows:

$$P(I_k \mid X_k) = \frac{P(X_k \mid I_k)\, P(I_k)}{P(X_k \mid I_k)\, P(I_k) + P(X_k \mid I_k^c)\, P(I_k^c)}. \tag{2}$$

Where we denote by $P(X_k \mid I_k^c)$ the conditional probability.

We now note that, if $P(X_k \mid I_k) = P(X_k \mid I_k^c)$ from (2) follows $P(I_k \mid X_k) = P(I_k)$, that is the information X_k does not modify the probability evaluation $P(I_k)$.

Generally, given the likelihood ratios

$$L_k = P(X_k \mid I_k)/P(X_k \mid I_k^c),$$

equation (2) becomes

$$P(I_k \mid X_k) = p\frac{L_k}{pL_k + (1-p)}. \tag{3}$$

By observing that $L_k/(pL_k + (1-p)L_k) = 1$, from the right-hand side of (3) we obtain, for the likelihood ratio L_k:

$$P(I_k \mid X_k) \geqslant P(I_k) \Leftrightarrow L_k \geqslant 1.$$

We could order C by means of a preference relation, ordering the concepts $c_1, c_2, \ldots, c_N$ according to the decreasing values of the conditional probabilities $P(I_1 \mid X_1)$, $P(I_2 \mid X_2), \ldots, P(I_N \mid X_N)$.

If we consider two concepts c_i, c_j, assuming that:

$$p_i' = P(I_i \mid X_i), \quad p_j' = P(I_j \mid X_j),$$

we obtain:

$$p_i' \geqslant p_j' \Leftrightarrow \frac{p_i'}{1-p_i'} \geqslant \frac{p_j'}{1-p_j'}. \tag{4}$$

Furthermore from (2) and (3) we obtain:

$$\frac{p_k'}{1-p_k'} = \frac{p}{1-p}L_k, \quad k = 1, 2, \ldots, N. \tag{5}$$

Thus expression (4) becomes:

$$p_i' \geqslant p_j' \Leftrightarrow L_i \geqslant L_j. \tag{6}$$

Therefore the ordering of the concepts in C according to decreasing values of the conditional probabilities $P(I_k \mid X_k)$, $k = 1, 2, \ldots, N$, can be obtained by ordering the ratios $P(I_k \mid X_k)/P(I_k^c \mid X_k)$ in the same way.

Since we assumed $P(I_k) = \text{const}$ for all k, this is equivalent to ordering the concepts in order of decreasing likelihood ratios L_k.

The computation of the quantities L_k, $k = 1, 2, \ldots, N$, requires evaluation of the probabilities $P(X_k \mid I_k)$, $P(X_k \mid I_k^c)$.

A feasible way of doing so can be by making the assumption that the components of the random vector $X = (x_1, x_2, \ldots, x_v)$ representing a concept in C are exchangeable.[11]

Then we make the assumption:

$$P(x_i = 1 \mid I) = a, \quad P(x_i = 1 \mid I^c) = b \quad \text{for } i = 1, 2, \ldots, v \tag{7}$$

where I is the event defined as: 'a concept randomly chosen in C is informative for the user'.

If we assume that the components of X are conditionally independent with respect to I and I^c, then from (7) follows (being $h = \sum_{i=1}^{v} x_i$),

$$P(X \mid I) = a^h(1-a)^{v-h}, \tag{8}$$
$$P(X \mid I^c) = b^h(1-b)^{v-h}. \tag{9}$$

Therefore $P(X \mid I)$ and $P(X \mid I^c)$ depend only on how many (and not which) components of the vector X are equal to 1, that is on how many and not which documents make references to the concept c. From the above follows that the exchangeability property holds, with respect to each of the two subsets of the informative and non-informative concepts.

From (8) and (9) follows:

$$\frac{P(X \mid I)}{P(X \mid I^c)} = L = \left(\frac{a}{b}\right)^h \left(\frac{1-a}{1-b}\right)^{v-h}. \tag{10}$$

The exchangeability property that leads to (10) also allows an ordering of C by giving 'informative weights' to the concepts.

From (10) and the assumption:

$$c' = \lg\frac{a}{b}, \quad c'' = \lg\frac{1-a}{1-b} \tag{11}$$

follows

$$\lg L = he' + (v-h)e'' = \mathscr{L}(e'-e'') + ve'' = w + e,$$

where $\quad w = h(e'-e''), \quad e = ve''.$

In (11) the 'total informative weight' of the concept c represented by X is given by w.

Thus, denoting by $w^{(i)}$ and $w^{(j)}$ the total weights of the concepts c_i and c_j, equation (6) becomes

$$p_i' \geqslant p_j' \Leftrightarrow w^{(i)} \geqslant w^{(j)}, \tag{12}$$

that is, if h_i and h_j are the numbers of documents of D containing respectively the concept c_i and c_j we get:

$$p_i' \geqslant p_j' \Leftrightarrow h_i \geqslant h_j \quad (\text{if } a \geqslant b),$$
or $$p_i' \geqslant p_j' \Leftrightarrow h_i \leqslant h_j \quad (\text{if } a \leqslant b).$$

That is, given two concepts, their preference order depends only on the number of documents that contain that concept.

Using this preference order, it is possible to order all the indexing terms which are related to the concept entered by the user.

The indexing terms are ordered by the system and the terms that are present at the top of the ordered set are presented to the user for browsing in the conceptual auxiliary data structure.

5. CONCLUSIONS

Making use of preference relations an associative algorithm has been formally justified. The importance of the algorithm is justified by its use into an hypertext interface in order to provide the user with a list of concepts associated to the user's one(s). In particular, the HYPERLINE prototype shows that the algorithm has viable applications even for very large databases.[21]

It has to be noted that the algorithm performs concept-to-concept associations at the time of the user interaction with the information retrieval system. This allows to implement hypertext interfaces that reply to the user input with lists of associated concepts in place of list of references.

In the HYPERLINE prototype the concept network where the associated concepts are identified has been built using available thesauri. Future work will be devoted to the identification of suitable automatic knowledge extraction and indexing tools to be used for the concept network creation. The associative algorithm described is independent from the characteristics of the IR system used (boolean or probabilistic), improvements

from a thigh integration with a probabilistic environment are matter for future work.

Acknowledgements

Pier Giorgio Marchetti wishes to thank Professor R. Scozzafava for the useful suggestions and discussions on preference relations applications.

The present work has been developed in the framework of a joint research activity carried on by the European Space Agency and the University of Padova.

The work of Maristella Agosti has been partly supported by the Italian National Research Council (CNR) under the project 'Sistemi informatici e calcolo parallelo – P5: Linea di Ricerca Coordinata MULTI-DATA'.

REFERENCES

1. M. Agosti, Is Hypertext a new model of information retrieval? IOLIM 1988, *Proc. 12th Int. Online Information Meeting*, vol. 1, pp. 57–62. Learned Information, Oxford, UK (1988).
2. M. Agosti, New potentiality of Hypertext systems in information retrieval operations. In *Human Aspects in Computing: Design and Use of Interactive Systems and Work with Terminals*, edited H.-J. Bullinger, pp. 317–321. Elsevier Science Publishers, Amsterdam, The Netherlands (1991).
3. M. Agosti, G. Gradenigo and P. Mattiello, The Hypertext as an effective information retrieval tool for the final user. *Pre-proceedings of the III International Conference on Logics, Informatics and Law*, vol. I, edited A. A. Martino, pp. 1–19. Firenze (1989).
4. M. Agosti, G. Gradenigo and P. G. Marchetti, Architecture and functions for a conceptual interface to very large online bibliographic collections. *Proc. RIAO 91, Barcelona, Spain*, pp. 2–24 (1991).
5. M. Agosti, G. Gradenigo and P. G. Marchetti, A Hypertext environment for interacting with large textual databases. *Information Processing and Management* **28** (2) (1992).
6. J. Aitchison and A. Gilchrist, *Thesaurus Construction – A Practical Manual*, 2nd edn. Aslib, London (1987).
7. P. D. Bruza, Hyperindices: a novel aid for searching in hypermedia. In ref. 30, pp. 109–122.
8. P. D. Bruza and Th.P. van der Weide, Assessing the quality of hypertext views. *SIGIR Forum* **24** (3), 6–25 (1990).
9. V. Bush, As we may think, *Atlantic Monthly* **176** (1), 101–108 (1945).
10. W. B. Croft and R. H. Thompson, I3R: a new approach to the design of document retrieval systems. *Journal of the American Society for Information Science* **38** (6), 389–404 (1987).
11. B. De Finetti, *Theory of probability* vols. 1 and 2. Wiley, New York (1975).
12. T. E. Doszkocs, Implementing an associative search interface in a large online bibliographic data base environment. New Trends in Documentation and Information, *Proc. of the 39th FID Congress, Edinburgh, UK*, pp. 295–297 (1978).
13. H. P. Edmundson and R. E. Wyllys, Automatic abstracting and indexing – survey and recommendations. *Communications of the ACM* **4**, (5), 226–234 (1961).
14. M. E. Frisse, Searching for information in a hypertext medical handbook. *Communications of the ACM* **31** (7), 880–886 (1988).
15. M. E. Frisse and S. B. Cousins, Information retrieval from hypertext: update on the dynamic medical handbook project. *Hypertext '89 Proc., Pittsburgh, Pennsylvania*, pp. 199–212 (1989).
16. P. Ingwersen and I. Wormell, Improved subject access, browsing and scanning mechanisms in modern online IR, *Proc. 9th ACM-SIGIR Conf. on Research and Development in Information Retrieval, Pisa, Italy*, pp. 68–76 (1986).
17. P. E. Jones, R. M. Curtis, V. E. Giuliano and M. E. Sherry, Application of statistical association techniques for the NASA document collection. *Technical Report NASA CR-1020*. Arthur D. Little Inc. (1968).
18. D. Lucarella, A model for hypertext-based information retrieval. In ref. 30, pp. 81–94.
19. H. P. Luhn, The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 159–165 (1958).
20. P. G. Marchetti and N. J. Belkin, Interactive online search formulation support. *National Online Meeting, New York* (1991).
21. P. G. Marchetti and G. Muehlhauser, HYPERLINE: the information browser. *Esa Bulletin* **66**, 115–118 (1991).
22. W. A. Martin, AUTOSEARCH: a proposed clustering procedure for the automatic searching of online information retrieval systems. *Proc. 7th National Online Meeting, New York* (1986).
23. J. Nielsen, The art of navigating through hypertext. *Communications of the ACM* **33** (3), 296–310 (1990).
24. S. E. Robertson and K. Spark Jones, Relevance weighting of search terms. *Journal of the American Society for Information Science*, 129–146 (1976).
25. S. E. Roberton, C. J. van Rijsbergen and M. F. Porter, Probabilistic models of indexing and searching. In *Information Retrieval Research*, edited R. N. Oddy, S. E. Robertson, C. J. van Rijsbergen and P. W. Williams, pp. 35–56. Butterworths (1981).
26. G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill (1983).
27. G. Salton and C. Buckley, On the use of spreading activation methods in automatic information retrieval. *Proc. 11th ACM-SIGIR Conf. on Research and Development in Information Retrieval, Grenoble, France*, pp. 147–160 (1988).
28. L. J. Savage, *The Foundations of Statistics*. Dover Publications, New York (1972).
29. B. Shneiderman and G. Kearsley, *Hypertext Hands-on! An Introduction to a New Way of Organizing and Accessing Information*. Addison-Wesley, Reading, MA (1989).
30. N. Streitz, A. Rizk and J. André (eds), Hypertext: concepts, systems and applications. *Proc. of the 1st European Conf. on Hypertext, INRIA, France*. Cambridge University Press, Cambridge (1990).
31. C. J. van Rijsbergen, A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation* **33**, 106–119 (1977).
32. C. J. van Rijsbergen, *Information Retrieval*, 2nd edn. Butterworths (1979).
33. C. J. van Rijsbergen, A new theoretical framework for information retrieval. *Proc. 9th ACM-SIGIR Conf. on Research and Development in Information Retrieval, Pisa, Italy*, pp. 194–200 (1986).
34. H. D. White, Toward automated search strategies. *Proc. 13th Int. Online Information Meeting, London, UK*, pp. 33–47 (1989).