A knowledge based approach for automatic Indexing and Information retrieval.

H. ALIANE[1], M.C. ROUAG[2], L. BOUKARI[2].
Basic Software Laboratory
Research Center on Scientific and Technical Information
Rue des 3 frères Aissou Ben Ak,oun Alger,
ALGERIA

Abstract: This paper presents some aspects of an ongoing research and development project that aims at elaborating a unified multilingual platform that will support automatic indexing, information retrieval and knowledge discovery.
The system's design is based on second generation technology principles. The core of the system is a semantic net that serves to represent documents and concepts involved in an information retrieval system.
Multilingual, friendly user interfaces are offered for both expert user who has the privileges for the creation and management of a semantic network knowledge base and for the end user whose view of the system is one of a navigation based hypertext.
Two ways are considered for indexing:

- The system is considered as a semantic network-based thesaurus and then full thesaurus indexing can be performed.
- Indexing is not full-thesaurus based, thus indexing is based on concept discovery from texts.

Key-words: automatic indexing, intelligent information retrieval, semantic nets, concept discovery, multilingual interfaces.
CSCC'99 Proceedings:- Pages 1261-1265

1 Introduction

In the field of large textual knowledge bases, correctness of text indexing and efficiency of response to a simple query are the two crucial needs.
In the early 1966, Salton recommended the automatisation of these tasks because the manual realization is expensive and non- deterministic [10]. As an example of manual indexing disadvantages two different indexers using the same thesaurus in 30% of the cases don't index with the same descriptors.
In addition, information that doesn't seem relevant to the indexer could be of importance [1].
On the other hand, problem specific information retrieval is not only complicate by the amount of information but additionally by a lack of structure [3]. In order to succeed in performing rapid and selective information retrieval, it is indispensable to develop information providers on the basis of conceptually structured knowledge. To achieve these objectives, there are several approaches to develop application independent representation systems and to define suitable application programming interfaces to enable the sharing of information.
Our aim in this work is to develop a unified knowledge representation platform that will support automatic indexing, information retrieval and knowledge discovery. In the design and development of our system, we proceed in two steps; the first step, builds in partnership with human expert a semantic net knowledge-based thesaurus, that is used to perform full thesaurus automatic indexing and retrieval using a statistical method.
In the second step, the semantic net-based thesaurus is considered as a bootstrap that initializes the network in the sense of Pitrat[9]. Thus, indexing is based on a concept discovery process.

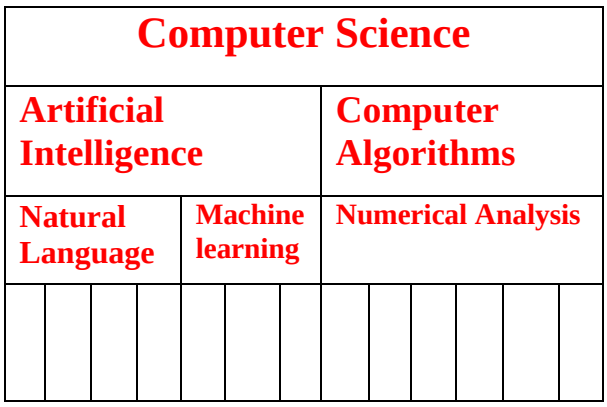

Fig.1

2 The semantic network-based thesaurus

Semantic nets: Among the knowledge representation schemes proposed in AI, semantic networks are often considered to be one of the most natural representations; indeed, with a network of interconnected links, the relationships between objects and patterns of interest can be represented explicitly. A semantic net is a structure for representing knowledge as a pattern of interconnected nodes and links [4]. Modern semantic nets are often credited to quillian's work on "semantic memory". Since then, many different versions have been implemented; although the terminology and notations may vary, the following characteristics are common to most of them :

- Nodes represent concepts of entities, attributes, events and states.
- Links represent conceptual relationships that hold between concept nodes. Labels on the links specify relationships types.

The system's kernel: the first feature of our system is a human-machine ensemble based design. "The human's role is primary to achieve total system performance objectives as a manager of knowledge resources that can vary in kind and amount of "intelligence" or "power" [3]. We aim at developing a complex information network that would be based on an existing thesaurus. In this step, highly visual interfaces are offered to a human expert to build his semantic net-based thesaurus. The system is domain independent in the sense that different experts could create their different knowledge bases. The thesaurus can be viewed as a large semantic network of terms (concepts) where links are of two types: relations between non-index and index terms and set-superset relationships (like is-a links). Fig.1 shows a portion of the semantic network. In order to expand the semantic structure of this information network, we incorporated actual documents and authors into the network. Documents are indexed by the various terms in the thesaurus, information is embedded in the document entry such as title, year of publication, call number, ...

We derived our semantic network-based online thesaurus by extracting a portion of the computer readable form of BN-OPALE subject-heading concerning the area of computer science.

This structure which constitutes the kernel of our system, will serve as the knowledge base for information retrieval.

We define nodes and links in our network more precisely as follow:

Nodes:

Nodes represent information about terms, documents or authors that are stored in the online thesaurus.

Terms : terms represent concepts. These terms may be used to index documents (index terms). Other synonymous terms are considered non index terms. Terms represent subject area knowledge , they constitute the knowledge base of the system.

Documents: they are the final outputs of the system. Typically, document nodes contain information about authors, year of publication, call number,…

Authors: searching for a particular author's publications is often used as a way to identify relevant works in the areas those authors are known in.

Links:

There are three kinds of links in our network: links for terms, subject-heading links and author links.

Links for terms: we consider the following links:

- The synonymous term link: leads a non-index term to its synonymous index term;
- The related term link: connects two related index terms;
- The neighboring term link: connects an index term to its neighbor term.

Heading links: connect index terms with document records. The relationship is many to many and bi-directional .

Author links: connect authors with document records. The relationship is also many to many and bi-directional.

These links enhance the semantic structure of our information network.

3 Indexing and text analysis

Early in the sixties, stils showed the importance of taking into account simultaneous occurrences of terms [12]. Controversially, other authors recommended linguistic methods for indexing. Several systems use lexical, syntactic and sometimes semantic knowledge. Other systems use hybrid statistical and linguistic approaches. However, linguistic approaches that seem at prior the most suitable (since they are based on text understanding) are very difficult to implement.

In order to address the indexing problem in our platform, we adopt the following postulate:

Frequency of events is significant

and then two approaches are proposed to perform indexing:

3.1 Full thesaurus-based indexing

This first approach exploits the knowledge-based thesaurus to index documents using the full thesaurus.

Our system uses a statistical method to compute concepts frequencies into documents, then, it performs a concept mapping algorithm to group concepts: Semantic relationships between concepts are used to map concepts given that each concept can be understood and located by its relation to other concepts. Main concepts are chosen to index documents on the basis of heuristic considerations that are to be confirmed by a real-life system.

3.2 Concept discovery

Nevertheless, the method described in 4.1 is no more useful when new concepts appear in the documents or when the thesaurus is incomplete or not exhaustive. Hence, our aim is the automatisation of concept selection as understood by AI researchers; the system may discover new concepts that do not exist in the knowledge base.

Our approach is based on concept discovery from texts and is inspired from [7]. The method is simple: we work with a minimal knowledge even incomplete. We do not use syntactic nor semantic analysis. In fact, current methods used in computational linguistics are not adequate (they are expensive to implement and it is not realistic to build a parser for each considered domain). "weak linguistic methods" could be more efficient [2]. Our concept discovery process is based on filtering texts by using linguistic schemas. The underlying ideas of using schemas is that: " studying co-occurrences of words and taking benefit of some regularities of language in the corpus especially in scientific and technical domains may lead to discover interesting patterns of information, hence new concepts"[2]. The semantic network is initialized by the bootstrap described above. The postulate is applied on the patterns discovered . The method can be described as follow (fig 2):

- A stop word list is used to remove non-semantic bearing words such as prepositions.
- A lexical analysis is performed to find known concepts i.e. concepts that already exist in the knowledge base and to find words with particular suffixes.
- The postulate is used to identify repetitive word sequences and to find configurations that may denote new concepts.
- Interesting configurations are identified using schemas: a configuration may denote a new concept if it corresponds to a schema in the schemas base.

We call "schema" an interesting linguistic configuration that may design a concept. Schemas constitute the schemas base. This base is built by the expert or induced by learning if the corpus is important.

Example of interesting configurations:

Unknown word	a word that specifies a schema	Known concept
Known concept	a word that specifies a schema	Unknown word

Other schemas are based on the repetition of some particular suffixes in the corpus, this will lead to induce significant concepts especially in scientific and technical domains for example "tion", "ie", "ome", … in medicine.

4 Retrieval

On the other hand, the semantic network serves as a navigation based hypertext that facilitates effective and efficient information retrieval. Indeed, Such an architecture encourages users to find information by browsing i.e. following a likely path from one node to another until they obtain their goal. This problem process is similar to a search in the problem space [5]. However, in the absence of effective search assistance, this method of search may require a lot of browsing and backtracking and can cause users to become lost in the search space; that's why effective search strategies must be implemented.

4.1 Search in the information network

Halasz indicated in [8], search and query in a hypertext network is one of the top issues of the next generation of hypertext systems, specifically he commented that:

"the note-cards experience suggests however, that navigational access by itself is not sufficient. Effective access to information stored in a hypermedia network requires query-based access to complement navigation"

Salton [11] presented a glimpse of a future hypertext system. He argued that vocabulary matching methods and link traversal algorithms (spreading activation) can be used to retrieve related parts of documents (corresponding to related nodes in a hypertext network). We believe the same argument can be applied to a hypertext that represents concepts and documents for information retrieval. A powerful search mechanism is necessary for alleviating the disorientation problem in a large information network. Therefore, our system uses a thesaurus activation strategy in addition to the classical vocabulary matching method.

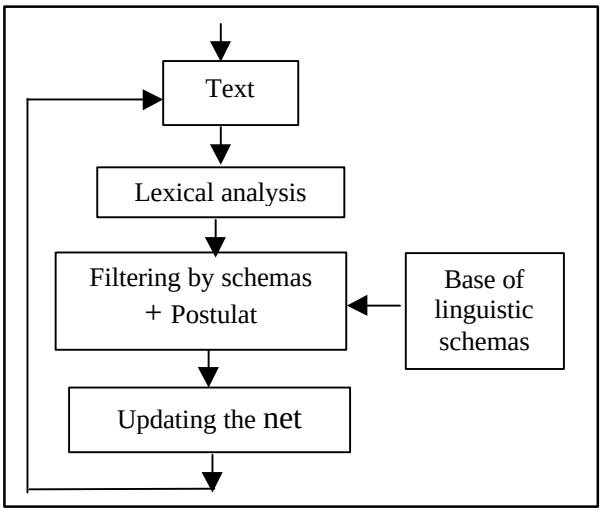

Fig.2

4.2 Thesaurus activation strategy

Thesaurus-activation strategy is often used by reference librarians to assist the user. In order to retrieve documents that would address the specific needs of the user; the librarian solicits search terms from the user. These user-supplied terms then have to be sharpened and translated into index terms [5]. The chance of terms in the user's query matching index terms is generally low. The librarian therefore initiates a terms translation process which includes consulting the thesaurus and a brainstorming process. The goal of this process is to identify the most specific terms to represent the query. In this stage of the query, both the user's and the librarian familiarity with the subject area plays an important role in determining the appropriate terms. The consultation terminates when a reasonable number of relevant documents is found. This strategy is appropriate for a query of the following nature [5]:

Find all documents that are related to the following concepts: expert systems, fifth generation language, …

When the vocabulary matching method fails in satisfying the user's query, this later may use the hypertext navigation alternative.

We developed a program to simulate the thesaurus-activation strategy. Our system first uses the terms supplied by the searchers that matches with some nodes in our online thesaurus (vocabulary matching). These nodes are taken as source nodes, with which there are associated links. Our system applies a specific-terms-first heuristic spreading activation process on the semantic network based online thesaurus to generate relevant terms.

The Specific-Terms -First Heuristic:

On analyzing the structure of BN-OPALE we observed that nodes which have fewer neighbors in the semantic network are generally more specific in content than those that have more neighbors. Our system applies a heuristic which expands the nodes with fewer neighbors first (i.e. the more specific terms).

A session of user relevance feedback then follows. This thesaurus activation process is useful in helping users to articulate their queries.

5 Multilingual interfaces

Another important point we deal with in the design of our platform is multilinguism. In fact, we aim at handling the following situation:

- The document collection may be multilingual, in our particular case, languages are French, English and Arabic.
- The end user may use any of the above languages.

Actually, the end user can choose one of the above languages to interact with the system via different interfaces: the choice is made at the beginning of a user session.

Concepts are represented in the semantic network in French as main concepts. Equivalent terms in English and Arabic are represented as related terms to the main concepts in order to retrieve all documents (in different languages) corresponding to a user request.

6 Conclusion

In information science, use of a knowledge base for "intelligent" information retrieval has drawn significant attention in recent years.

We presented in this paper a knowledge based approach for automatic indexing and information retrieval. The system is about being evaluated by real world input.

Our information network represents the essential knowledge elements for information retrieval. We have chosen to found the platform on human-machine ensemble design because we believe the role of information specialists may reduce search indeterminism and ensure the validity of the kernel information input. This is important since this kernel will serve as a base or will initialize concept discovery and learning algorithms. For retrieval purposes , a thesaurus activation process generates relevant terms by activating the links in the semantic network based online thesaurus. This process is made possible by the use of specific-terms-first heuristic algorithm. On the other hand two ways of indexing are proposed: a full thesaurus based indexing using a statistical and concept mapping method. The second way reposes on a concept discovery process. Actually, we're about enhancing the concept mapping method and the

base of linguistic schemas. Other heuristic search strategies are also being implemented.

References:

[1] Andreewsky A, Fluhr C, "indexation automatique- construction automatique de thesaurus- classification automatique",. *Note CEA-N1795.*

[2] Aliane H. " notes on knowledge acquisition using linguistic schemas", *research report*, 1997.

[3] Bursner S, Spreckelsen C. " from pure information handling to knowledge management in medicine" position paper, *second knowledge engineering forum*, feb 96.

[4] H. Chen, K. Basu and T, Ng "An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): symbolic branch and bound search vs connectionist Hopfied Net Activation", *MIS Research report*, 1994 .

[5] H. Chen, "A knowledge -based design for hypertext-based document retrieval systems" *proceeding DEXA* 90.

[6] H. Chen, B. Schatz, J. Martinez, T. Ng "Generating a domain-specific Thesaurus Automatically: An experiment on FlyBase.", *MIS research report*, 1994.

[7] C. Enguehard, P. Malvache, P. Trigano "Indexation de textes : l'apprentissage des concepts", *proceeding coling* 92, 1992.

[8] F.G. Halasz. Reflections on note-cards: seven issues for the next generation of hypermedia systems. *Communications of the ACM, 31(7):836-852, July 1988.*

[9] Pitrat J, *"textes, ordinateurs et comprehension"*, eyrolles, 1985

[10] Salton G. "information dissemination and automatic information systems", *proc, IEEE*, 54, 12, december, 1966.

[11] Salton G, Buckely C. " Parallel text search methods ", *communications of the ACM*, 31(2),1988, pp 202-215.

[12] Stiles H. F "the association factor in information retrieval", *journal of the ACM*, vol. 8, 1961, pp 271-279.

[13] Woods D, Roth E. M, "*cognitive systems engineering*" in Helander M (ed): hanbook of human-computer interaction . Elsevier Science publishers.