

Deriving Concepts Hierarchy

Yousef ABUZIR
AlQuds Open University
Jerusalem, Palestine
yabuzir@qou.edu

Abstract. Information Retrieval (IR) covers the problems relating to the effective storage, access, searching and locating documents that are relevant for user's information need or query from large collection documents. Many techniques and tools have been developed to improve these processes. One of these tools is the thesaurus. This paper will present a tool for users to build thesauri according to their own requirements. The objective of this paper is to present an overview of how thesauri can be created automatically. The main goal is to design a comprehensive thesaurus-building tool that can be used in any specialty field. The system provides sufficient structure to represent different relationships among concepts and terms as well as to extract these relationships. In addition a thesaurus builder can also use different methods to extract terms and relation between them from document corpus. So, the system provides a set of interface functions through which the user can easily build a thesaurus.

1 INTRODUCTION

A thesaurus is a set of concepts in which each concept is represented with at least synonymous terms, broader concepts, narrower concepts, and related concepts. A term is a word or sequence of words that refers to a concept [1]. The relations or links broader, narrower, related, and synonymous have been defined for thesauri [2, 3]. Originally intended for indexing and retrieving documents, thesauri have increasingly been seen as knowledge bases and used beyond the domain of librarianship.

Thesaurus has long been a concern in lexicography, and recently, it has found many applications in machine translation, information retrieving, and computational lexical semantics, etc. [4,5,6]. A thesaurus is a structured system of terms encoding explicitly semantic relations like synonyms, hyponyms, hyperonyms, part-whole

relations, associations, etc. Because manual construction, maintenance, and updating is highly time-consuming, in the recent years methods have been developed that automatically construct a thesaurus out of a collection of documents related to a given subject. These methods are based on statistical or linguistic treatment of the document collections. Within this context the task of extracting terminological terms out of a text arises. Thesauri constructed in this way are in general less structured.

A thesaurus is a valuable tool in Information Retrieval (IR), both in the indexing process and in the searching process, used as a controlled vocabulary and as a means for expanding or altering queries. Most thesauri that users encounter are manually constructed by domain experts and/or experts at document description. Manual thesaurus construction is a time-consuming and quite expensive process, and the results are bound to be more or less subjective since the person creating the thesaurus make choices that affect the structure of the thesaurus. There is a need for methods of automatically construct thesauri, which besides from the improvements in time and cost aspects can result in more objective thesauri that are easier to update. Therefore, building thesauri automatically from corpus received a large attention in recent years.

In the second section of this paper, we summarize the research issues involved in the general problem examined in this paper, as well as the solutions proposed for several of these issues. The problem of thesaurus construction can be broken down to the following sub-problems term extraction and term organization - deriving relationships between terms from texts. Section 3 presents a general description of the Thesaurus Construction Tools and the techniques used. Sections 4 and 5 present experimental results of the different models for term extraction and building the hierarchical relationships

between terms. Section 6 discusses the conclusions and future work.

2 THESAURUS CONSTRUCTION

The most conventional, approach to thesaurus construction is to build it manually. This labor intensive and hence only possible in specialized domains where repeated use may justify the cost. This approach requires a domain expert to prepare a hierarchical structure of topics relevant to a particular subject area. Searchers then employ terms from this hierarchy to form queries that automatically expanded to more specific or general terms.

The growing amount of information and the need for quick access to it slowly changes the significance of Information Retrieval (IR) systems. In order to turn Information Retrieval systems into more useful tools for both the professional and general user, one usually tends to enrich them with more intelligence by integrating information structure, such as thesauri. Since it is difficult and expensive to build thesauri manually, many researchers attempted to construct thesauri automatically. This section, discusses the different approaches to thesaurus construction. It also describes some systematic procedures for thesaurus construction.

The construction of thesauri generally takes the following form:

Extract word co-occurrences.

Define similarities (distances) between words on the basis of cooccurrences.

Cluster words on the basis of similarities.

2.1 The Different Approaches

There are different approaches to construct thesaurus automatically. Figure 1 shows these different approaches. We can classify these approaches into two groups. In the first one we can group the different approaches according to the source of terms used to construct the thesaurus. The later is based on the corpus-based techniques.

2.1.1 Approaches Based on Source of Data

There are three approaches to construct a thesaurus based on the soured of terms. The first approach, on designing a thesaurus from

document collection, is a standard one [7, 8]. Here the idea is to use a collection of documents as the source for thesaurus construction. This assumes that a representative body of text is available. By applying **statistical** or **linguistic** procedures we can identify important terms as well as their significant relationships.

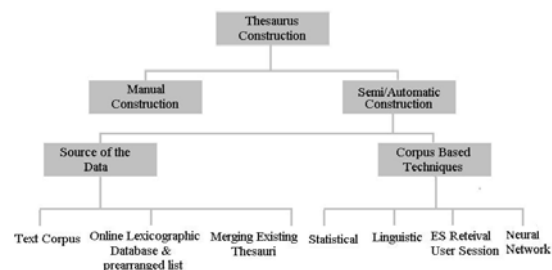


Figure 1 The Different Approaches to Thesaurus Construction.

The simplest approach is to reuse existing online Lexicographic database, such as wordNet or Longman's subject codes [9].

Pre-arranged list of terms like table of content, index of book or other format can be used to construct thesaurus automatically [10].

The third approach is merging existing thesauri. This third approach is appropriate when two or more thesauri for a given subject exist, that need to be merged into a single unit [11, 12].

2.1.2 Approaches Based on Collection Treatment

In the second group, the classification based on the different corpus techniques used to build thesauri automatically. A corpus-based method performs a computation on the text of the documents in the corpus to induce a thesaurus.

Subsumption - construct a hierarchical thesaurus from a computed list of complex noun phrases where subsumption roughly corresponds to the subset relation defined on terms (i.e.; "intelligence" subsumes "artificial intelligence") [13].

Linguistic - Several researchers have used head-modifier relationships to determine semantic closeness [14, 15]. Co-occurrence statistics and head-modifier relationships get at different kinds of information. Two terms that modify the same words (or are modified by the

same words) often belong to the same semantic category.

Statistical - approaches based on semantic relatedness [16] by considering the occurrence of terms in document. Documents are clustered into small groups based on a similarity measure that considers two documents similar if they share a significant number of terms, with medium frequency terms preferentially weighted. Terms are then grouped by their occurrence in these document clusters. Other statistical approaches construct thesauri by transposing the standard term-by-document matrix in order to define a similarity measure on terms rather than documents [7, 17].

Expert System -The third automatic is based on tools from expert systems [18]. In this approach thesauri are built using information obtained from users. For example, if a retrieval system user combines two terms by OR, in his/her query these terms are probably synonyms.

Neural Network - The fourth approach is based on Self-Organizing Map (SOM). SOM is an unsupervised two-layer neural network used to summarize large high-dimensional data. Roussinov and Chen explained how they used Kohonen's SOM as a tool for extracting semantic relationships between words and creating a hierarchy of categories [19]. Their prototype system creates a hierarchy of concepts from document returned by a WWW search engine using SOM.

3 AN OVERVIEW OF THE SYSEM

This section will present our system, which is a thesaurus construction system that can be used as a tool to build domain independent thesaurus.

3.1 General

Our system is a tool for users to build thesauri according to their own requirements. The main goal is to design a comprehensive thesaurus-building tool that can be used in any speciality field. It provides sufficient structure to represent different relationships among concepts and terms as well as to extract these relationships. In addition a thesaurus builder can also use different methods to extract terms and relation between them from document corpus. So, the system

provides a set of interface functions through which the user can easily build a thesaurus.

The system provides two different modes depending on the user. These two modes are training and expert mode. In the first, users can learn and experiment the different methods for automatic term extraction and finding relationship between these terms. The second one can be used or assisted expert who is interesting in building domain independent thesaurus.

The system allows easy to extract the required terms. It also provides the ideal environment to build the hierarchical relations between these terms. It contains different Models that can be used or assisted in thesaurus construction (see Figure 2):

Term Selection Models – the extraction of terms (indexing terms) from text document. Automatic term extraction approaches like Document frequency (DF), Inverse like Document frequency (IDF), Discrimination Value (DV) and Frequency of Occurrence are used to find index terms. These statistical models predict the strengths of terms from the frequencies of these terms in the collection.

Terms co-occurrence Models - the occurrence of two terms or more can be perceived as an evidence of a relation between these words. The text documents are read in word by word. Whenever one of the keywords occurred with the word in query, the sentences are displayed and statistical data are stored to find the strength of the relation between terms. Term co-occurrence used to produce structures or maps of related terms.

Hierarchical organization of terms Models – Models for building hierarchical relationships between terms from text. Three models support creation of hierarchical relation are used, the first model that produce concept structure with an ordering from general terms to more specific that use cohesion statistic to measure the degree of association between terms. The second one is use the lexico-syntactic Analyzer to find the relationship between terms. The other one is use the subsumption.

HTML Hierarchical Model - to build a hierarchical structure of terms automatically from a set of HTML documents.

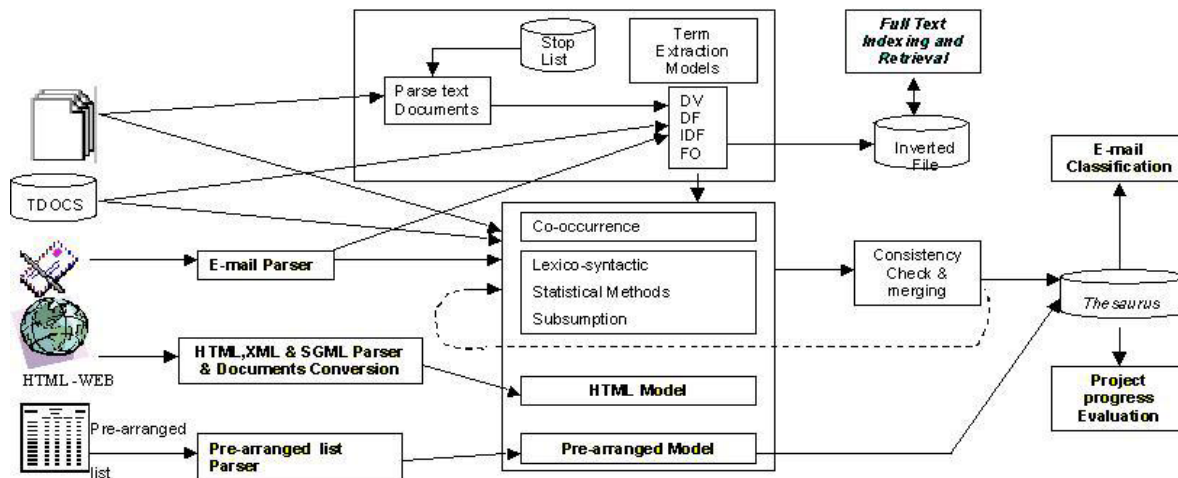


Figure - 2 Architecture of the System

ThesConv Model – Thesaurus Converter has import function to convert different input formats like book's index to TDOCS [20] Thesaurus Manager Data Base and spreadsheet. The result database is powerful for processing thesaurus relations.

Full search and retrieval Model - This model retrieves the required documents from the collection. A list of documents names can be obtained by giving search term(s). User can browse the selected documents.

Documents Conversion Models – This Model converts HTML, SGM and XML document to ASCII format. Other application can extract email messages and does information extraction from the email messages. Sender, subject, etc, are extracted from the email message.

3.2 Term Extraction

The objective of this tool was to automatically build domain independent thesaurus from a set of documents. This translates into three basic phases: term extraction, relationships deriving and filtering and reviewing of the terms and relationships.

The problem of thesaurus construction can be broken down into the following sub-problems:

- Term extraction.
- Term organization - deriving relationships between terms from texts.

Term extraction: terms for the thesaurus were to be extracted from the documents and had to best reflect the topics covered within them.

Relationships deriving: these models provide the user with the relationships between terms

The final stage is the filtering and reviewing of these relationships. Before converting these results into the database of the thesaurus the user had to remove the noisy relationships that are in the result.

Some terms extracted from document text by Term Extraction Model may not function effectively as indexing terms in a thesaurus. A decision must be made about what terms should be included in the thesaurus. To decide what to include, various methods like Document Frequency (DF), Inverse Document Frequency (IDF), Discrimination Value (DV), and Frequency of Occurrence are used.

4 Lexico-syntactic Model

In order to extract terms and relations between terms, different methods can be used. One of these methods requires predefinition of lexico-syntactic patterns as well as manual traverses on outputs of terminologists, in order to find pairs that belong to the predefined relations. Hearst [21, 22] reports a method using lexico-syntactic patterns to extract lexical relations.

The patterns were constructed during manual analysis of text documents in the area of Total Quality Management (TQM). A sample of 48 different documents relates to Total Quality Management was used. Some patterns may simply not be applicable to a particular Broader-

Narrow Term (BT-NT), while in other cases there may be a natural default pattern.

A lexico-syntactic expression is composed of a set of elements, which can be lemmas,

Pattern	Pair of term (BT-NT) Relationships	A Sample Example
for	(system, organization) (system, Management system) (process, Project) (document, record)	It is a system for translating the strategic intent on an organization to the managerial and operating spheres of decision making.
As	(system, organization) (system, process) (supplier, service)	He views the organization as a system and advocates using a scientific method to optimize the system

The experiment was meant to find positive evidence of various patterns. The system uses one pair of terms at a time to extract sentences that contain this pair. For each pair a list of sentences where the pair occurred in the text was made. From these lists common patterns were identified. In some cases a particular patterns included more than concepts (pair of terms), as can see in Table 1. Table 1 includes some examples of the patterns. The pattern formats extracted in our experiment are primitive. An example, in the first case we tried (system, organization) pair, and with just this pair we found new patterns. Next we tried other pair and discovered more patterns.

By using the Document Viewer of the system all keywords terms are directly visible. The system combines ideas of term identification, relationship extraction for terms and label these terms and relationship. The system requires only few manual definitions of lexico-syntactic pattern and the need to know these lexico-syntactic patterns in advance.

At the third step, a set of sentences is extracted. These sentences are lemmatized, and terms are identified. So, we represent a sentence by a lexico-syntactic expression. For example, the relation (*system, management system*) allows extracting the following sentence from the corpus TQM:

“However, in their study of target analysis system at Nissan, Balachandran and Srinidhi (1991) found the use of target analysis more as a management system to manage not only costs but also quality.”

From this sentence, we produce the lexico-syntactic expression: Term1 as Term2 but also Term3.

punctuation marks, or words, such as TERM1, etc. Through this simplification process, we have a more generic representation of relevant sentences, and comparing these sentences is easier.

The following sentences (Table 2) extracted from TQM corpus show other examples of the hyponym relation between terms:

<i>Thus, <u>suppliers include business such as distributors, dealers, warranty repair service, transportation, contractors, and franchises, as well as that provides materials and components.</u></i>	Term1 such as Term2, {Term _i } and Term _n
<i><u>Supplier also include service suppliers, such as health care, training, and education providers.</u></i>	Term1 also include Term2

The system supports interactive extraction of the terms and relations using the Lexico-Syntactic Model. With just the previous Lexico-syntactic patterns we find more new patterns as well as new terms (Figure 3). Note that for these patterns, even though they have emphatic term, this does not affect the fact that the relation indicated is hyponymic. The application of this technique to TQM domain showed great success.

5 Deriving Concept hierarchies from HTML

5.1 Introduction

All the approaches to thesaurus construction mentioned so far perform what we might call

construction by content, since information for constructing a thesaurus is extracted from the text of the document. Thesaurus construction by content does not exploit an essential aspect of a hypertext environment like the Web, namely the structure of documents and the link topology. In this section we investigate a technique for automatic thesaurus construction, which we have called construction by context, since it exploits the context in the HTML document to extract useful information for building thesaurus. Thesaurus construction by context exploits relevance hints that are present in the structure and topology of the HTML documents published on the Web. Combining a large number of such hints, an adequate degree of accuracy can be achieved in constructing the thesaurus.

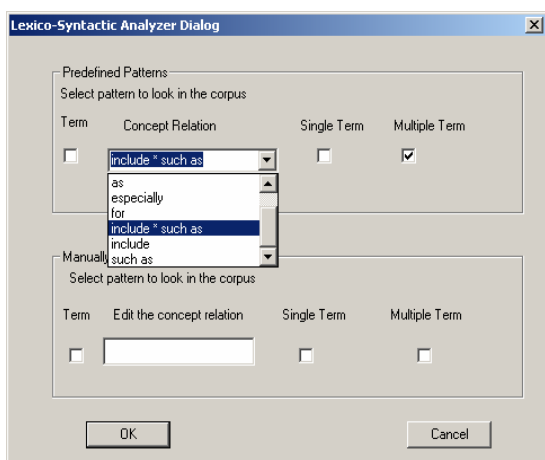


Figure 3. Lexico-Syntactic Dialog

The focus of this section is the construction and evaluation of an automatic tool for thesaurus construction from HTML pages. The system is a tool which, given a specific URL address of the web page that is broad and well-represented of the domain, will find out and return a list of terms that it considers the most authoritative for that topic. It is built to perform a local analysis of text, meta-tags and links to arrive at a "global consensus" of the best terms and relationships between them for specific domain.

5.2 System Overview

Figure 4 shows an overview of the construction process, which is made up of several steps. First the application accepts URL address as input and the HTML page is fetch.

From this HTML document plain text, meta-data like meta-tags, titles, headings, list, etc and URL are extracted. The outcome is stored in a database and used in later stages.

For each HTML document, the tool parses the text and classifying it into a maximum of three different groups. The groups give a local analysis on text, met-data and links found in the HTML page. These groups are (see figure 5):

- Metadata: as present in HTML META-tags and important text, like document title and all HTML headings.
- Plain text: all other text.
- URL list: the list of the URL in the HTML page.

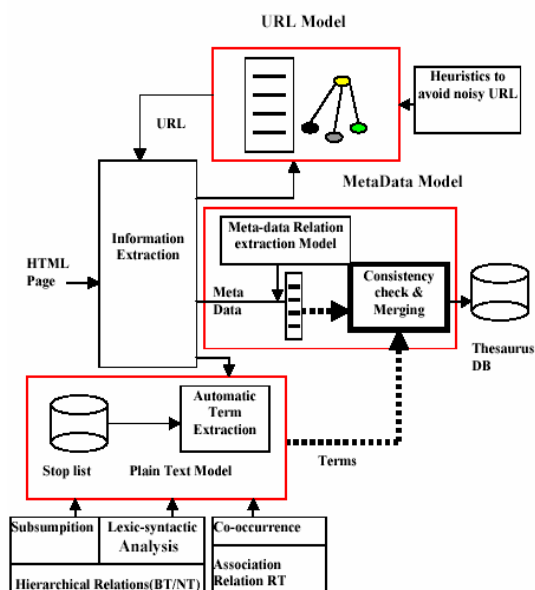


Figure 4 Architecture of the HTML Model

5.3 URL Extraction Model

Any URL found during the parsing is passed to processing if it points to a document within the current site, and stored for later analysis if it points to an external site, currently limited to depth of 3. This allows performing a depth-first visit of a site, collecting any categorization information it contains about itself and other sites.

Quite often the HTML pages of a site have a characteristic structure represented by links across pages. For instance there can be references to the main page, or links to the general services available in the site, like searching within the

site, help or information pages. Finally, there can be advertisement banners in precise positions in each page. We want to avoid processing such pages. In order to identify these structural links, we used some heuristics about the structure of URL and an initial analysis of pages reachable from the starting page. These links placed in a stop list of URLs and discarded in the subsequent analysis of the site.

This task starts from a list of URLs, retrieving the documents referred by each of them and analyzing the structure of the document expressed in terms of its HTML tags.

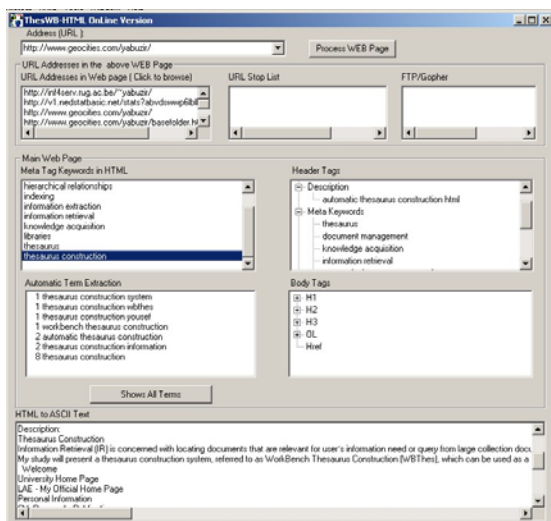


Figure 5 Architecture of the HTML Model

5.4 Meta-Tags Analysis

The tags considered are currently <TITLE>, <Hn>, , , <A>, etc. and a meta-tags <META Classification Keywords>. Whenever one of these tags is found, a context phrase is recorded, which consists of the title within a pair <Hn> </Hn>, or the first portion of text after a or tag, or the phrase within a <A> tag.

In the analysis, tags related to layout or emphasis (, <I>, <CENTER>, etc.) are discarded here. Such tags can be effectively used in plain text analysis to extract thesaurus terms.

In general the system will build the structure of the web pages as Follows:

1. Accept a URL, i.e. HTML document.

2. The system defines a target node using a start tag and an end tag, start tag is used to label each node.
3. Define a target low-level node within the node.
4. Repeat step 3 if more low-level nodes to be defined.
5. Repeat steps 2-4 if more nodes to be defined.

In order to extract the relationship between terms, the system exploits the hierarchical structure and sets of extraction rules. For each node in the structure, it is associated with rules that extract that particular information or element.

During the parsing process, the system applies text extraction rules for each type of tags. There are rules to extract and build the tree structure of the tags <Hn>. Such a tree has the tag <TITLE> as a root. The extraction rule for each of these tags is applied until all tags have been extracted.

If a tag is a list, the extraction rule for the list will be applied iteratively to extract the list elements for that node in the structure. After that, other extraction rules will be applied to extract individual items like terms, compound terms, lexico-syntactic patterns, looking for subsumption, etc. As a final step, we use the structure to group together the individual items to assemble the thesaurus.

When we build the structure from an HTML document, the page is first split to a sequence of nodes. The extraction rules for each of the nodes in the structure are applied to the inner of tags until all tags have been consumed. After that, more rules are used to extract individual items.

6 Conclusion and future Work

We described a system for building a thesaurus from electronic documents and present results based on different combinations of techniques. The system serves as basis for a generalized framework for automated thesaurus construction tool.

The system is a semi-automatic thesaurus construction tool. It makes it easy to construct thesaurus by automating many of the tasks. The tool can be used for training new user on how to build the thesaurus. At the same time expert can use the tool to build his or her own thesaurus.

Different models for automatic thesaurus construction are presented in this work. Getting hierarchical relationships between terms entails an important advance in the process of thesaurus construction.

We addressed the problem how to effectively use the rich information, which is typically available in HTML to build thesaurus. We specified relevant hints about the hyperlinks and meta-data about web sites, which we hope, will aid in the task of thesaurus construction and aim at exploiting the richness of HTML. We examined these hints with different collection of web pages. We hope that our analysis will help future research into hypertext by providing some ideas about various types of information that may be present in a hypertext corpus and how one should take advantage of each.

The tool has the power for extracting the different sets of candidate terms and discovering the relationships between these terms. Although, the results achieved with our tool are quite encouraging. In most cases, it achieves an effective and accurate result. The tool needs manual constraints and the different models do not prevent from extraction of pairs which are not link by the target relation.

REFERENCES

1. R. Rada and B. Martin Augmenting Thesauri for Information Systems, *ACM Transactions on Office Information Systems*, 5 (4), 378--392 (1987).
2. ISO 5964, Documentation - Guidelines for the Development and Establishment of Multilingual Thesauri, 1985.
3. J. Aitchison, A. Gilchrist, and D. Bawden, *Thesaurus Construction*. 3ed. ASLIB, 1997.
4. J. Klavans, and E. Tzoukermann 'Combining corpus and machine-readable dictionary data for building bilingual lexicons' In *Machine Translation*, Vol.10, No.3, pp 1--34,1996.
5. M. A. Hearst, and H. Schutze, "Customizing a Lexicon to Better Suit a Computational Task", *Proceedings of 31st Annual Meeting of ACL*, Columbus, Ohio, USA, 55-69, 1993
6. P. Resnik, "Disambiguating Noun Groupings with respect to WordNet Senses", in S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann and D. Yarowsky (eds.), *Natural Language Processing Using Very Large Corpora*. Kluwer Academic Publishers, 1999, pp. 77-98. (Slightly revised version; the original paper appeared in the *Proceedings of the 3rd Workshop on Very Large Corpora*, MIT, 30 June 1995.
7. G. Salton, M. J. McGill, *Introduction to modern information retrieval*. McGraw Hill, New York. 1983.
- 8 Y. Abuzir, D. Vervenne, P. Kaczmarek and F. Vandamme "E-mail messages classification and user profiling by the use of semantic thesauri", *Proceeding of CIDE2001 - 4e Colloque International sur le Document Électronique*, Toulouse - France 24-26 octobre 2001.
9. E. M. Voorhees, "Using WordNet to Disambiguate Word Sense for Text Retrieval", *Proc ACM SIGIR'93*, Pittsburgh, 1993, 171-180.
- 10 Y. Abuzir "ThesConv: A tool for thesaurus construction from Prearranged list." to appear in proceedings of the 2002 International Conference on Information and Knowledge Engineering (IKE'02) part of a International Multiconferences in Computer Science, Monte Carlo Resort, Las Vegas, Nevada, USA, June 24 - 27, 2002.
- 11 H. Mili, R. Rada "Merging Thesauri: Principles and Evaluation". *IEEE Transactions On Pattern Analysis and Machine Intelligence*,10(2):204-220, 1988.
12. M.V. Mannino, S. B. Navathe, and W. "A Effelsberg , Rule-Based Approach for Merging Generalization Hierarchies". *Information Systems*, 13(3):257-272, 1988.
13. D. A. Evans, W. K. Ginther, M. Hart, R. G. Lefferts, and Monarch, I. A., Automatic indexing using selective nlp and first-order thesauri, In *Proceedings of the RIAO*, vol. 2, pp. 624-643, 1991.
14. G. Grefenstette, Use of syntatic context to produce term association lists for text retrieval. In *SIGIR'92*, pp. 89--97, 1992.
15. G. Ruge, Experiments on linguistically based term associations. In *RIAO'91*, pp. 528-545, 1991.
16. C. J. Crouch, An approach to the automatic construction of global thesauri, *Information Processing & Management*, 26(5): 629-40, 1990.
17. Y. Qiu, H.P. Frei, Concept Based Query Expansion. *Proc. of the 16th Int. ACM SIGIR Conf. on R&D in Information Retrieval*, Pittsburgh, SIGIR Forum, ACM Press, June 1993.
18. Güntzer, U., Jüttner, G., Seegmüller, G. and Sarre, F., Automatic Thesaurus Construction by Machine Learning from Retrieval Sessions. *Information Processing & Management*, Vol. 25, No. 3, pp. 265-273, 1989.
19. D. Roussinov and H. Chen, "A scalable self-organizing map algorithm for textual classification: A neural network approach to thesaurus generation." *Communication Cognition and Artificial Intelligence*, 1998.
20. Y. Abuzir, D. Vervenne, P. Kaczmarek and F. Vandamme, "TDOCS Thesauri for Concept-Based Document Retrieval", *R&D Report BIKIT, BIKIT - LAE*, University of Ghent 1999.
21. Hearst M. A., Automatic Acquisition of Hyponyms from Large Text Corpora. In *Actes, 14th International Conference on Computational Linguistics (COLING'92)*, pages 539-545, Nantes, France, 1992.
22. Y. Abuzir, D. Vervenne, P. Kaczmarek and F. Vandamme, "Extracting Semantic Relationships between Terms using IKEM Tool", *KIM/KIT NEWS*, Vol. 15, nr.3, Nov. 2000.