COENOSES 13(2): 89-98, 1998

INTERACTIVE THESAURUS CONSTRUCTION METHODS IN THE ECOLOGICAL DOMAIN

F. Maggiore[1] & C. Anzaldi[2]

1 Istituto per lo Studio della Dinamica delle Grandi Masse-Consiglio Nazionale delle Ricerche, San Polo 1364, 30125 Venezia, Italy, Tel +39 (041) 5216882, Fax +39 (041) 2602340, e-mail: maggiore@flux.isdgm.ve.cnr.it
2 Istituto Applicazioni del Calcolo-Consiglio Nazionale delle Ricerche, Viale del Policlinico 137, 00161 Roma, Italy, Tel.+39 (06) 88470271, Fax +39 (06) 4404306

Keywords: Information retrieval; Language processing; Syntactic methods; Text analysis; Terminology.

Abstract: Researchers having to deal with enormous amounts of information gleaned from international data banks are familiar with the benefits of a specialized, well-constructed thesaurus for information retrieval. In the ecology field considerable amounts of specialized, interdisciplinary information can be obtained from international data banks that are continuously updated. We describe here the construction of an interactive thesaurus for information retrieval in a subdomain of the ecology, specifically the ecology of the Venice Lagoon. The thesaurus was constructed using a multilingual bibliography (Italian, French and English texts). Two approaches to information retrieval are used: 1) examination of text by grammatical analysis, leading to the extraction of significant phrases; and 2) semantic clustering of these phrases using the word root. The design is described by applying it to a subset of an example bibliography.

Introduction

A specific thesaurus has been constructed to organize a knowledge-based bibliography of the ecology of the Venice lagoon (Anzaldi & Maggiore 1994). The aim of the present work is to summarize the methodology of construction using a subset of the bibliography. There has been a recent revival of interest in thesauri as reference tools for information retrieval. The thesaurus can be viewed as a filter for an information array that grows very rapidly. Such a filter is a powerful tool if the thesaurus is specialized and customized to the end user.

According to the ANSI definition (Guidelines 1978), a thesaurus in a given domain is defined as a hierarchical structured set in which the components exhaustively describe a given domain. These components are known as terms, the morphology of which must correspond to several precise rules. There exist many international thesauri based on accepted disciplinary knowledge (International Energy Subject Thesaurus 1993; Thesaurus for the Environment 1995). Why then did we construct a thesaurus on an interdisciplinary field, managed directly by the user and constantly updated? First, because environmental research is becoming increasingly interdisciplinary (Anzaldi et al. 1996); second, the information problem is not merely one of retrieval, but of selection (from the Internet, for example) so that the pertinent domain needs to be delimited with care by the researcher; third, exponential increases in research information have resulted in the need to continuously update scientific language. A thesaurus should therefore be easy to consult and to update.

Traditional thesauri cannot completely satisfy any one of these exigencies because of the fundamental principle that they must be disciplinary-wise complete and very large. They therefore require so much time to prepare that, when ready for publication, they are already in part obsolete. Moreover, they are strictly dependent on the language of the country that constructed them, so for internationalization they also require time for translation. Also, some technical terms are not well translated from the original language (e.g. 'raton', the Spanish translation for 'mouse', is not used because it also has a vulgar significance). Updating an accepted, traditional thesaurus has the same problems as its construction. By contrast, a thesaurus in a well-specified domain constructed with the advice of experts can be very helpful.

From the viewpoint of information retrieval, researchers having to deal with enormous amounts of information gleaned from international data banks are aware of the benefit of a specialized, well-constructed thesaurus. In constructing such a thesaurus, two main purposes are distinguished: the selection of appropriate items from a very large database, and a greater depth of knowledge in the selected set of items. In the first case, the thesaurus operates well as a qualified filter; in the second case, it operates as a powerful tool for information retrieval on a specific archive.

The methodology described here, which takes into account the above observations, incorporates two steps:

(1) Extraction of keywords (KW) from the text of scientific subject bibliographies in a specific field. The methodology performs an analysis of text to achieve a meaningful list of KWs, which are gathered in a list. The thesaurus, reasonably constrained and easily consulted, is built up in a hierarchical structure to more levels. These methodologies are supported through active interaction with an expert in the field.

(2) Indexing and retrieval of text using the thesaurus KWs.

While such systems are domain-independent, they work best with accurately specified and restricted domains. Applications of the system are established by researchers for their own bibliographies and search projects, but also include the organization of scientific journal searches.

Methods of Thesaurus Construction

In the past few years, many specialists have attempted to resolve problems in thesaurus construction by borrowing from automated or semi-automated approaches in computer technology, such as information retrieval and natural language processing (e.g. Drewes 1983; Crouch 1990; Blosserville et al. 1992; Wang et al. 1985; Heinrichs 1992; Salton et al. 1990). There are two major challenges in creating and maintaining a thesaurus: updating the information base, and maximizing the ability of the information retrieval system. Some experts suggest a bottom-up approach to updating a thesaurus. In this approach, one begins with the terms currently used in research, teaching and business, and extracts from these a sector terminology that can be used to build up and readily update the existing thesaurus. To maximize the efficiency of information retrieval and achieve multi-language capability, lexical roots can be used since most technical terms are coined from Latin or Greek roots. (Nencioni 1994), allowing for the ready identification of semantically related terms.

Thesaurus construction is based on the extraction of terms (syntagmas) from texts using syntactic and semantic analysis methods, as well as on hierarchic structuring via clustering based on word roots. The rigour of the scientific language used in a domain leads to language being exemplified by means of a specific nomenclature, repeated throughout the text, which facilitates the identification of significant syntagmas. As some authors (Drewes 1983; Crouch 1990; Blosserville et al. 1992) have underscored, the application domain definition has the important effect of introducing a specialist language in which words with a wide-ranging, general meaning are restricted to a circumscribed area of knowledge, thereby allowing a considerable reduction in ambiguity.

For a thesaurus constructed in this way, two major applications are envisioned:

(1) as a terminological tool for a very close description of the domain;

(2) as a retrieval tool.

From the viewpoint of retrieval, two main purposes are distinguished: selection from a very large database (e.g. Internet querying), and deepening of knowledge in the set of selected, specific items. In the first case the thesaurus works as a qualified filter, while in the second it works as a powerful retrieval tool for a specific archive.

Methods of Information Retrieval

The growing quantity of information received in response to a given query is becoming increasingly difficult to manage. Methodologies of information retrieval that analyze the complete text result in a very wide and thorough search, but require the use of powerful tools and sophisticated techniques that often give results of high recall but low accuracy (Salton et al. 1990). In this work, an a priori choice has been made to 'privilege' recall in the search with respect to accuracy (use of the terms 'privilege' and 'accuracy' follows Salton 1968). Heinrichs (1992) proposed that a set of interest should be extracted from large information databases, and that this subsets should then be examined further using detailed queries at the local level.

In view of the above, it was deemed of interest to develop a system for local use in which the query was closely linked to the thesaurus and its construction. When a filter is required to select information from many very large databases, it is useful to construct an interface that utilizes the thesaurus automatically. When the thesaurus is used for information retrieval, the system presented here allows for the automatic indexing of archives using the thesaurus, and for Boolean retrieval on indexed archives. We introduced this utility since automatic indexing is occasional whereas retrieval is performed numerous times.

The method works in two distinct phases:

(1) Indexing of texts using the thesaurus KWs;

(2) Querying of the text using the same KWs.

The completeness and relevance of the thesaurus is thus tested by means of a judgment based on both the indexing obtained and the recall and precision of retrieval. The presentation is supported by the guided construction of a thesaurus coming from a subset (Anzaldi and Maggiore, 1994).

The system works on texts, in particular abstracts and titles. In preliminary trials, the algorithm applied to titles and to abstracts gave the same results. Therefore, taking into account that the title of a scientific article generally summarizes the objects treated, it is often considered sufficient to utilize titles in a first approximation. This approach is also useful in achieving efficiency in elaboration and management. Thus the bibliography on the ecology of the Venice Lagoon is processed using article titles (Anzaldi & Maggiore 1994).

Construction of the Thesaurus

The thesaurus is constructed by applying syntactic and semantic methods to the contents of texts dealing with an accurately defined scientific domain. Of course, the initial selection of the texts is extremely important. Here we present

Table 1. Titles of scientific texts.

Balanus amphitrite (Cirripedia, thoracica), a potential indicator of fluoride, copper, lead, chromium and mercury.
Fluoride accumulation in barnacles and mussels.
Effect of PCBs in *Leander adspersus*: bioaccumulation, oxygen consumption, osmoregulation.
Hydrocarbon uptake and loss by the mussel *Mytilus edulis*.
Levels of mercury, cadmium and lead in *Serranus cabrilla* and *Serranus scriba*.
Gonadal steroidogenesis in teleost fish.
Steroid biosynthesis by the ovary of the European eel, *Anguilla anguilla* L.
Spatial distribution of plankton community along a salinity gradient in the Venice Lagoon.
Phytoplankton concentration in the Malamocco Channel of the lagoon of Venice.

the methodology applied to a subset of the entire archive (Anzaldi and Maggiore 1994).

Syntactic Analysis

The method proposed uses syntactic analysis to extract *phrases* from texts and to obtain terms by means of a progressive, partially interactive process of refinement of these *phrases*. The analysis is based on three lexica: lexicon of function words, negative lexicon, and lexicon of genitive terms (only the preposition 'of' in the English language). An example list of titles to which the methodologies are applied is presented in Table 1.

The genesis of the *phrases* occurs by applying the following rules:

(1) Identification of separating elements that cut text into fragments of autonomous, meaningful *phrases*. Taking into account the rules of syntactical phrase determination, parentheses, apexes, targets of punctuation, conjugations and prepositions (with the exception of the genitive) have been considered as separating elements (lexicon of function words). The lexicon of function words is used to break up the texts in order to obtain *phrases* that we will denote KEYPHRASES. These have a grammatical structure of varying complexity but all have an independent meaning.

(2) Normalization of the KEYPHRASES. This operation includes phrase reduction using the negative lexicon to eliminate redundancy, as well as interactive refinement. The negative lexicon comprises all the general items of a language: definite and indefinite articles, numerals, possessives, demonstrative or indefinite adjectives and pronouns, mono-composed KEYPHRASES of one or two characters, genders and ways of nouns, adjectives and pronouns, verbal declinations and composed prepositions. KEYPHRASES produced by applying a lexicon of functional words and a negative lexicon to the titles in Table 1 are given in Table 2 (column A).

A third lexicon is used to resolve ambiguities related to the use of the genitive. The preposition "of" cannot always be used as a separator. For example "indicator of fluoride" has a different semantic value than "indicator" and "fluoride" separately; the meaning of the compound term being the ability of an organism to evidence the presence of fluoride in the environment; the meaning of "indicator" and "fluoride" being more general. From the point of view of the domain, the first term falls in the domain of Water Quality, the second is very generic, while the third falls in the domain of chemical elements. In the case presented, "of" cannot be considered a separator. By contrast, "problem of ecological genetics" can be broken down to "problem" and "ecological genetics", because the significant KEYPHRASE is "ecological genetics"; "problem" is too generic and should be discarded. KEYPHRASES produced by applying the genitive are presented in Table 2 (column B). In this way KEYPHRASES that constitute meaningful elements are produced from the initial text. The list of KEYPHRASES obtained represents a first sketch of thesaurus.

Semantic Analysis

In constructing a thesaurus, it is necessary to increase the accuracy of term formation and to introduce relations between terms and identifying synonyms. A number of solutions to this resolution problem have been proposed (e.g. Drewes 1983; Wang et al. 1985; Salton et al. 1990; Anick et al. 1993), but the inadequacy of these methods is recognized. Since text analysis requires specific competence in the domain, and since it is dependent on the cultural background of the thesaurus builder, it is necessarily subjective and cannot be treated by mere processing, however complex that processing may be.

In the present work, the kind and form of terms and the degree of specificity are established a priori by the expert. To obtain a non-structured thesaurus formed by terms, an analysis of the meaning of each KEYPHRASE is required. In setting up a specialist domain thesaurus that can run on a personal computer, direct knowledge storage has been replaced by an expert's direct contribution. Using all his knowledge, the expert interacts with the computer in dealing with the problems of semantic analysis as they occur. The semantic analysis is then applied to the KEYPHRASES (Table 2, columns A and B).

A number of potential problems are recognized:

Table 2. Column A: KEYPHRASES produced by applying lexicon of function words and negative lexicon to the titles of Table 1; Column B: KEYPHRASES produced by applying lexicon of genitive; Column C: TERMS produced by interaction on the KEYPHRASES.

COLUMN A	COLUMN B	COLUMN C
BALANUS AMPHITRITE		BALANUS AMPHITRITE
CIRRIPEDIA		CIRRIPEDIA
THORACICA		
POTENTIAL INDICATOR OF FLUORIDE	POTENTIAL INDICATOR FLUORIDE	INDICATOR OF FLUORIDE
COPPER		INDICATOR OF COPPER
LEAD		INDICATOR OF LEAD
CHROMIUM		INDICATOR OF CHROMIUM
MERCURY		INDICATOR OF MERCURY
FLUORIDE ACCUMULATION		FLUORIDE ACCUMULATION
BARNACLES		BARNACLE
MUSSELS		MUSSEL
EFFECT OF PCBs	EFFECT PCBs	EFFECT OF PCBs
LEANDER ADSPERSUS		LEANDER ADSPERSUS
BIOACCUMULATION		BIOACCUMULATION
OXYGEN CONSUMPTION		OXYGEN CONSUMPTION
OSMOREGULATION		OSMOREGULATION
HYDROCARBON UPTAKE		HYDROCARBON UPTAKE
LOSS		HYDROCARBON LOSS
MUSSEL MYTILUS EDULIS		MUSSEL MYTILUS EDULIS
LEVELS OF MERCURY	LEVELS MERCURY	LEVEL OF MERCURY
CADMIUM		LEVEL OF CADMIUM
LEAD		LEVEL OF LEAD
SERRANUS CABRILLA		SERRANUS CABRILLA
SERRANUS SCRIBA		SERRANUS SCRIBA
GONADAL STEROIDOGENESIS		GONAD STEROIDOGENESIS
TELEOST FISH		TELEOST FISH
STEROID BIOSYNTHESIS		STEROID BIOSYNTHESIS
OVARY OF EUROPEAN EEL	OVARY EUROPEAN EEL	OVARY EEL
ANGUILLA ANGUILLA		ANGUILLA ANGUILLA
SPATIAL DISTRIBUTION OF PLANKTON COMMUNITY	SPATIAL DISTRIBUTION PLANKTON COMMUNITY	SPATIA L DISTRIBUTION OF PLANKTON
SALINITY GRADIENT		SALINITY GRADIENT
VENICE LAGOON		VENICE LAGOON
PHYTOPLANKTON CONCENTRATION		PHYTOPLANKTON
MALAMOCCO CHANNEL OF LAGOON OF VENICE		MALAMOCCO CHANNEL
	LAGOON VENICE	LAGOON OF VENICE

(1) "and", comma, "or", etc. could separate one or more elements referring to one or more identifying components. In the KEYPHRASES "potential indicator of fluoride", "copper", "lead", "chromium", "mercury"; "hydrocarbon uptake" and "loss", the identifying components are "indicator", "uptake", and "loss". The first refers to the biological indicators of water quality, while the other two refer to the ability of organisms to release or take up pollutants. In this example, every element should be reassigned to such components to provide additional KEYPHRASES to ensure that each has an exhaustive meaning and contributes equally to the retrieval process. By contrast, the KEYPHRASES *Serranus cabrilla* and *Serranus*

scriba maintain a self-sufficient meaning (Table 2, column C), since they denote specific fish species.

(2) A semantic analysis is carried out concerning the role of the genitive as a separator element; dividing KEYPHRASES and analyzing interactively the parts, the subdivision is not accepted when the genitive adds semantic value. For an expert user looking for information on the effects of pollutants on organisms, the KEYPHRASE "effects of PCB" has a meaning inherent in a specific query that would not be answered by the single terms "effects" and "PCB" (PolyChloro-Byphenils). The first term has no scientific meaning if kept isolated, while the second differentiates a chemical compound. In such cases the KEYPHRASE should not be decomposed.

(3) In the list of the KEYPHRASES (Table 2, columns A and B), changes are carried out based on decisions reflecting representativeness and specificity:

(a) KEYPHRASES without purely scientific meaning, or the parts of KEYPHRASES that make them redundant (e.g. "potential") are excluded; also, those that are too specific for a given context are excluded. Such decisions are made at the level of expert consultant based on the specified purpose of the thesaurus.

(b) KEYPHRASES can be broken down (where necessary and possible, without changing the meaning) into constituent terms; thus "gonadal steroidogenesis" can be divided into "gonadal", and then modified into "gonad" and "steroidogenesis."

At the end of this phase a list of scientific TERMS refined to the degree of required specificity is obtained (Table 2, column C). This may be referred to as a "non-structured thesaurus". The next step is to structure the thesaurus into a hierarchical tree, by joining terms in KEYWORDS of more general meaning and recognizing synonyms. This step requires application of a root methodology. The structuring of a set of terms has attracted the interest of several authors (e.g. Crouch 1990; Wang et al. 1985; Paice & Jones 1993), all of whom emphasize the improved query efficiency afforded by a structured thesaurus (Guidelines, 1978). In structuring a set of terms, a methodology based on the consideration that a root identifies the semantic field of a family of words is adopted. The root of the words making up each term is therefore extracted, and a possible semantic clustering is suggested. In technical and scientific fields, this operation is facilitated by the widespread use of Greek words and the tendency of Greek to favour noun compositions (Nencioni 1994). The use of roots allows for the joining of terms of related meaning. For example "bioaccumulation" and "fluoride accumulation" are joined under the root ACCUM. The terms refer to bioaccumulation of substances in living organisms; INDICA joins five terms referred to the ability of an organism to signal the presence of heavy metals (mercury, lead, etc.) in the environment (Table 3). The roots so identified represent meaningful structures that can be utilized for the identification of KEYWORDS of more general terms. In the aforementioned examples, the KWs will be respectively BIOACCUMULATION and BIOLOGICAL INDICATORS (Table 3, column C). The root may also brings to light synonymies between one or more terms. For example, the root STEROI joins the synonymies "steroidogenesis" and "steroid biosynthesis" (Table 3, column B) to form a parallel structure or synonymic phrase.

Since the thesaurus will be used to retrieve information from a specific ecological field, it should be representative of the field and establish a uniform scientific level. Some sorting and recognition are therefore required. In identifying roots and derivative terms, the following problems may arise:

(1) A screening may be necessary to exclude word roots that are too specific, such as species epithets (e.g. adsper, amphit, cabrill, edulis, scriba),. Generic terms and words identifying generic topics are also screened out (Table 3).

(2) For terms composed of two or more words, a decision can be made regarding the number of corresponding roots to retain. This decision depends on a priori choices made regarding thesaurus structure; for example SPATIAL DISTRIBUTION OF PLANKTON has the roots SPATIA, DISTRI, PLANKT; from this, two KWs (SPATIAL DISTRIBUTION and PLANKTON) can be produced, or it may be decided to select only PLANKTON.

To maximize the domain definition, synonyms can be introduced. This operation can only be carried out with an interactive analysis in which the user has specific knowledge of the topics. Similarly, the conversion from roots to KWs must be carried out by an expert. Because the system operates as the expert system that can increase its original knowledge, thesaurus updating requires less resources than its construction.

The expert interactively makes the correct choices and assigns a keyword to each subgroup, the meaning of which includes the generalized concepts contained in the terms comprising the subgroup. Finally, the expert aggregates the keywords into larger subgroups and identifies these subgroups with a topic. The result is a two-level structured set that may be considered a thesaurus. In the implementation described there are three levels of processing.

Indexing and Querying Methods

The thesaurus, whose construction we have described, is ready to be used as a terminological tool for the domain. It can also be used from the system for the purpose of indexing and querying archives. In the present paper, it was decided to index the archive in order to perform interactive querying. The subdivision of the two operations into different phases, one carried out only once and the other presumably quite frequently, is consistent with the stated aims (restricted and specialized domain; structured and standardized terminology; processing on small computers). Indeed, the first phase is time consuming because each keyword is searched in the

ROOT	Column A	Column B	Column C
	TERMS	SYNONYMS	KEYWORDS
accumu	bioaccumulation		bioaccumulation
accumu	fluoride accumulation		bioaccumulation
adsper	Leander adspersus		
amphit	Balanus amphitrite		
anguil	Anguilla anguilla	eel	Anguilla anguilla
balanu	Balanus amphitrite		Balanus amphitrite
barnac	barnacle		Crustacea
bioacc	bioaccumulation		
biosyn	steroid biosynthesis		
cabril	Serranus cabrilla		
cadmiu	level of cadmium		
channe	Malamocco Channel		
chromi	indicator of chromium		
cirrip	Cirripedia		Crustacea
consum	oxygen consumption		oxygen consumption
copper	indicator of copper		
distri	spatial distribution of plankton		spatial distribution
edulis	Mytilus edulis		
eel	eel		
effect	effect of PCB		
fish	fish		Pisces
fluori	fluoride accumulation		
fluori	indicator of fluoride		
gonad	gonad		gonads
gradie	salinity gradient		
hydroc	hydrocarbon loss		
hydroc	hydrocarbon uptake		
indica	indicator of chromium		biological indicators
indica	indicator of copper		biological indicators
indica	indicator of fluoride		biological indicators
indica	indicator of lead		biological indicators
indica	indicator of mercury		biological indicators
lagoon	Venice Lagoon		
lagoon	lagoon of Venice	Venice Lagoon	lagoons
lead	indicator of lead		
lead	level of lead		
Leander	Leander adespersus		Leander adespersus
level	level of mercury		bioaccumulation
level	level of cadmium		bioaccumulation
level	level of lead		bioaccumulation
loss	hydrocarbon loss		hydrocarbon loss
malamo	Malamocco channel		lagoon of Venice
mercur	indicator of mercury		
mercur	level of mercury		
mussel	mussel		Mollusca
mytilu	Mytilus edulis		Mytilus edulis
osmore	osmoregulation		osmoregulation
ovary	ovary		gonads
oxyge	oxygen consumption		
PCB	effect of PCB		
phytop	phytoplankton		
plankt	spatial distribution of plankton		plankton
plankt	phytoplankton		plankton
salini	salinity gradient		physical properties
scriba	Serranus scriba		
serran	Serranus cabrilla		Serranus cabrilla
serran	Serranus scriba		Serranus scriba
spatia	spatial distribution of plankton		
steroi	steroidogenesis	steroid biosynthesis	biosynthesis
steroi	steroid biosynthesis		
teleos	teleost		Pisces
uptake	hydrocarbon uptake		hydrocarbon uptake
Venice	lagoon of Venice	Venice Lagoon	lagoon of Venice
Venice	Venice Lagoon		

Table 3 (opposite page). Roots extracted from terms of Table 2, column C. In columns A, B and C, root terms are in alphabetical order. Empty spaces in column C indicate that the corresponding root has been discarded.

Table 4 (below). An example of the thesaurus.

Topics	Keywords	Terms	Synonyms
Animal Organs	Gonads	Gonad	
		Ovary	
Biochemistry	Biosynthesis	Steroidogenesis	Steroid Biosynthesis
Communities Studies	Plankton	Phytoplankton	
		Plankton	
		Spatial Distribution of Plankton	
Geographic Areas	Lagoon of Venice	Lagoon of Venice	Venice Lagoon
		Malamocco Channel	
Habitat	Lagoons	Lagoon of Venice	Venice Lagoon
Oceanography	Physical Properties	Salinity Gradient	
Physiology	Bioaccumulation	Bioaccumulation	
		Fluoride accumulation	
		Level of Cadmium	
		Level of Lead	
		Level of Mercury	
	Hydrocarbon Uptake	Hydrocarbon Uptake	
	Hydrocarbon Loss	Hydrocarbon Loss	
	Osmoregulation	Osmoregulation	
	Oxygen Consumption	Oxygen Consumption	
Species	Anguilla anguilla	Anguilla anguilla	Eel
	Balanus amphitrite	Balanus amphitrite	
	Leander adspersus	Leander adspersus	
	Mytilus edulis	Mytilus edulis	
	Serranus cabrilla	Serranus cabrilla	
	Serranus scriba	Serranus scriba	
Taxa	Crustacea	Barnacle	
		Cirripedia	
	Mollusca	Mussel	
	Pisces	Fish	
		Teleost	
Water Quality	Biological Indicators	Indicator of Chromium	
		Indicator of Copper	
		Indicator of Fluoride	
		Indicator of Lead	
		Indicator of Mercury	
	Effects on Organisms	Effect of PCBs	

whole archive. The attribution of keywords performed in this phase makes the retrieval phase much quicker.

The part of the system concerned with automatic indexing is divided into two phases: (1) searching for terms (and synonyms) in the texts; (2) assigning relevant KWs to the article. These operations are performed within the context of thesauri construction principles. The indexed subset of Table 1 is summarized in Table 5.

After the articles have been indexed it is possible to perform a two-level search. The higher level (topics) is used to orient the user's choices. It is possible to request a topic and then to examine all the KWs linked to it in the thesaurus. The same method can be used to display the terms aggregated to

Table 5. An example of archive indexing.

1) *Balanus amphitrite* (Cirripedia, thoracica), a potential indicator of fluoride, copper, lead, chromium and mercury.
***BALANUS AMPHITRITE*, BIOLOGICAL INDICATOR, CRUSTACEA**
2) Fluoride accumulation in barnacles and mussels.
BIOACCUMULATION, CRUSTACEA, MOLLUSCA
3) Effect of PCBs in *Leander adspersus*: bioaccumulation, oxygen consumption, osmoregulation.
BIOACCUMULATION, EFFECTS ON ORGANISMS, *LEANDER ADSPERSUS*, OSMOREGULATION, OXYGEN CONSUMPTION
4) Hydrocarbon uptake and loss by the mussel *Mytilus edulis*.
HYDROCARBON LOSS, HYDROCARBON UPTAKE, MOLLUSCA, MYTILUS EDULIS
5) Levels of mercury, cadmium and lead in *Serranus cabrilla* and *Serranus scriba*.
BIOACCUMULATION, SERRANUS CABRILLA, SERRANUS SCRIBA
6) Gonadal steroidogenesis in teleost fish.
BIOSYNTHESIS, GONADS, PISCES
7) Steroid biosynthesis by the ovary of the European eel, *Anguilla anguilla* L.
***ANGUILLA ANGUILLA*, BIOSYNTHESIS, GONADS**
8) Spatial distribution of plankton community along a salinity gradient in the Venice Lagoon.
LAGOONS, LAGOON OF VENICE, PHYSICAL PROPERTIES, PLANKTON, SPATIAL DISTRIBUTION
9) Phytoplankton concentration in the Malamocco Channel of the lagoon of Venice.
307. **LAGOONS, LAGOON OF VENICE, PLANKTON**

each KW. The system uses the technique of Boolean searching in the user's query. The search, initially performed on the keywords, can be taken to a deeper level using the terms linked to them. The system selects the articles containing all the keywords requested; when searching for journal articles using terms, it selects those that contain even only one of the terms requested.

The following is an example of retrieval of information concerning bioaccumulation in mollusks. The topics relevant to our Thesaurus are PHYSIOLOGY and TAXA. The requested topics are therefore PHYSIOLOGY, TAXA.

KEYWORDS REFERRING TO THE REQUESTED TOPICS:

BIOACCUMULATION, OSMOREGULATION, OXYGEN CONSUMPTION, HYDROCARBON LOSS , HYDROCARBON UPTAKE, CRUSTACEA, PISCES, MOLLUSCA.

On the basis of this display the keywords are chosen. On supplying the KW BIOACCUMULATION to the system the articles containing BIOACCUMULATION will be selected:

Keywords requested = **BIOACCUMULATION**

Articles selected for the Keyword requested:

- Fluoride accumulation in barnacle and mussel
- Effects of PCBs in *Leander adspersus*: bioaccumulation, oxygen consumption, osmoregulation
- Levels of mercury, cadmium, and lead in *Serranus cabrilla* and *Serranus scriba*

If the aim is to restrict the search to bioaccumulation in mollusks, the query is implemented using two keywords:

Keywords requested = **BIOACCUMULATION, MOLLUSCA**

Article selected for the Keywords requested:

- Fluoride accumulation in barnacle and mussel

If more specific information on bioaccumulation is required, the terms aggregated to **BIOACCUMULATION** are displayed.

TERMS REFERRING TO THE KW REQUESTED:

BIOACCUMULATION, FLUORIDE ACCUMULATION, LEVEL OF CADMIUM, LEVEL OF LEAD, LEVEL OF MERCURY

The query is implemented using the terms FLUORIDE ACCUMULATION and Level of cadmium and the system selects all the articles containing FLUORIDE ACCUMULATION as well as all those containing Level of cadmium.

Terms: **FLUORIDE ACCUMULATION, LEVEL OF CADMIUM**

Titles selected for the terms requested:

- Fluoride accumulation in barnacle and mussel
- Level of mercury, cadmium and lead in *Serranus cabrilla* and *Serranus scriba*

Conclusions

The thesaurus has the following characteristics: (1) it has a bottom-up structure; (2) the terms are very close to the chosen domain; (3) the structure based on semantic hierarchy.

The proposed system can be used in two ways:

(1) Use of the thesaurus-constructing phase for terminological purposes (Anzaldi 1994). The use of this sort of thesaurus has very wide applicability:

- It is a storage tool of structured, normalized, and updated terminological information for a specific domain. The variability over time of scientific terminology leads to the replacement of obsolete or disused terms with terms having the same meaning but which are more rigorous and concise. This has led to the problem of uniformity of language. The introduction of synonymic phrases has allowed a reasonably good level of terminology standardization to be achieved.
- It utilizes a rational and objective search of a very large database, by implementing an automatic interface and using the Thesaurus as a complete, powerful filter.
- The system allows references to be ordered chronologically, making it possible to keep abreast of new developments in thinking and in scientific language.

(2) Use to follow the entire path of thesaurus building up, indexing and querying.

- It can be utilized for indexing archives in the construction domain and for retrieving through the thesaurus keywords. This itinerary (thesaurus building, indexing and retrieving) allows individual researchers to utilize the system for personal bibliographic purposes, or according to the above mentioned proposal of Heinrichs, for research projects; and by libraries intending to carry out research on journal articles; and for acquiring archives in machine-readable form from specialist journals or INTERNET data banks.

One interesting application of this method was the thesaurus on Energy and Environment for ENEA (Italian Agency for New Technology, Energy and Environment) for bibliographic and terminological purposes (Anzaldi and Bordoni, 1995).

Acknowledgement: The authors thank Jane Frankenfield Zanin for her useful suggestions on the manuscript.

References

Anzaldi, C. 1994. Un metodo para contribuir a la puesta al dia de la terminologia tecnica. IV simposio Iberoamericano de terminologia. Buenos Aires, pp. 37-42.

Anzaldi C. & L. Bordoni. 1995. Una applicazione del sistema SBIC: Costruzione di un Thesaurus nel settore Energia- Ambiente. Atti del Congresso Annuale AICA, pp. 385-389.

Anzaldi, C., L. Bordoni & A. Sano. 1996. Construction of a terminological interdisciplinary thesaurus. In: Galinsky C. & K. Schimtz (eds.). TKE'96: Terminology and Knowledge Engineering. Vienna, pp. 273-278.

Anzaldi, C. & F. Maggiore. 1994. Sistema di Gestione Bibliografica e Thesaurus su Biologia ed Ecologia della Laguna di Venezia. Istituto Studio Dinamica Grandi Masse-CNR, Venezia, TR 192.

Blosserville M. J., G. Hebrail, M. G. Monteil & N. Penot. 1992. Automatic Document Classification: Natural Language Processing, Statistical Analysis and Expert System Techniques used together. Internat. SIGIR Conference. Copenhagen, pp. 51-57.

Crouch, C. J. 1990. An approach to the automatic construction of global thesauri. Inform. Process. Manage. 26: 629-640.

Drewes, B. 1983. Retrieval of abstracts by analogy. Lectures notes in computer sciences, pp. 238-250.

Guidelines for Thesaurus Construction, Structure and Use. ANSI. 1978.

Heinrichs, N. 1992. The growing crisis of traditional information retrieval systems - what is to follow?. Lecture Notes in Computer Science 146: 1-12.

International Energy Subject Thesaurus. 1993. IEA.

Nencioni, G. 1994. Linguistica e terminologia tecnico scientifica. Quaderni del Lessico. Intellettuale Europeo. Lexicon Philosoficum, Roma 7: 5-12.

Paice, C. D. & P. A. Jones. 1993. The identification of important concepts in highly structured technical papers. Proc. of ACM SIGIR Conf., Pittsburgh, pp. 69-79.

Salton, G. 1968. Automatic Information Organization and Retrieval. McGraw Hill.

Salton, G., C. Buckley & M. Smith. 1990. On the application of syntactic methodologies in automatic text analysis. Inform. Process. Manage. 26: 73-92.

Thesaurus for the Environment. 1995. Trilingual version for Italy, Consiglio Nazionale delle Ricerche (ed.).

Wang, Y., J. Wanderdorpe & M. Evens. 1985. Relational thesauri in information retrieval. J. Am. Soc. Info. Sci. 36: 15-27.