Constructing and maintaining knowledge organization tools: a symbolic approach.

Fidelia Ibekwe-Sanjuan

Journal of Documentation, Emerald Publishing Group, 62 (2), 2006, 229-250.

Constructing and maintaining knowledge organization tools. A symbolic approach

Keywords. Knowledge organization, Thesaurus construction, Shallow NLP, Semantic relations acquisition, Term clustering, Information visualization.

Purpose
To propose a comprehensive and semi-automatic method for constructing or updating knowledge organization tools such as thesaurus.
Methodology
We propose a comprehensive methodology for thesaurus construction and maintenance combining shallow NLP with a clustering algorithm and an information visualization interface. The resulting system TermWatch, extracts terms from a text collection, mines semantic relations between them using complementary linguistic approaches and clusters terms using these semantic relations. The clusters are mapped onto a 2D using an integrated visualization tool.
Findings
The clusters formed exhibit the different relations necessary to populate a thesaurus or an ontology: synonymy, generic/specific and relatedness. The clusters represent, for a given term, its closest neighbours in terms of semantic relations.
Practical implications
This could change the way in which information professionals (librarians and documentalists) undertake knowledge organization tasks. TermWatch can be useful either as a starting point for grasping the conceptual organization of knowledge in a huge text collection without having to read the texts, then actually serving as a suggestive tool for populating different hierarchies of a thesaurus or an ontology because its clusters are based on semantic relations.
Originality of the paper
This lies in several points : combined use of linguistic relations with an adapted clustering algorithm which is scalable and can handle sparse data. We propose a comprehensive approach to semantic relations acquisition whereas existing studies often use one or two approaches. The domain knowledge maps produced by the system represents an added advantage over existing approaches to automatic thesaurus construction in that clusters are formed using semantic relations between domain terms. Thus while offering a meaningful synthesis of the information contained in the original corpus through clustering, the results can be used for knowledge organization tasks (thesaurus building and ontology population) The system also constitutes a platform for performing several knowledge-oriented tasks like science and technology watch, textmining, query refinement.
Category: Research paper

Introduction

As more and more texts continue to be produced in huge quantities everyday both on the Internet and through specialized communication forums (journals, books), the problem of updating knowledge repositories in order to keep up with their continual evolution becomes problematic. Thesauri, taxonomies or ontologies are well known structures for organizing and managing knowledge in different fields. In this paper, we will address specifically the case of thesaurus construction as being the typical knowledge organization tool useful for both professionals (documentalists and librarians) and information seekers. The problem of thesaurus construction and maintenance subsumes that of taxonomy as the latter is a simpler form of knowledge representation. Traditionally, thesauri were manually constructed (Aitchison, Ghilchrist & Bawden, 2000). However, this solution although more satisfying in terms of quality, is less feasible today. Manual construction and maintenance is a resource demanding and time-consuming task. Moreover, the results are rapidly overtaken in the present context of ever growing data on the information highway. A lot of research effort has been directed towards automatic or semi-automatic methods

for thesaurus construction (Rees-Potter, 1989 ; Grefenstette, 1994 ; Morin & Jacquemin, 2004 ; Schneider & Börland, 2004).
The existing methods fall under two main approaches : statistical approach and linguistic approach. None of the approaches used alone is sufficient to solve the many problems posed by automatic thesaurus construction, namely:

i. the automatic selection of domain terms,
ii. the automatic identification of thesaural relations between domain terms (generic/specific ; synonymy, relatedness, notes,..),
iii. the automatic construction of the actual thesaurus hierarchy and the horizontal relations (see also),
iv. maintenance of the thesaurus through acquisition of new terms, relations and incorporation into the thesaurus.

We propose an approach to thesaurus construction and update which brings answers to the first two problems (i-ii) and proposes a knowledge structure for steps (iii-iv) based on semantic clustering. This knowledge structure can be used either as a starting point or as an updating device for an existing thesaurus. Our approach combines linguistic and data analysis techniques in order to extract terms from a representative text collection, structure them through several semantic relations and cluster them into semantically coherent classes which can serve as a basis for thesaurus construction. As such, we propose an automatic assistance tool which can be placed within the category of "semi-automatic methods" for thesaurus construction. A domain specialist will have to validate the clusters produced and decide ultimately how to position the elements in the thesaurus. According to Schneider & Börland (2004) "*thesauri are fundamentally linguistic and conceptual in nature*[1]. *Structural, semantic and terminological problems are ever present, and manual intellectual construction work is necessary when dealing with these problems*". In agreement with this statement, we think that a more linguistically-founded approach is more suitable for automatic approaches to thesaurus construction.

With regard to earlier published works (Ibekwe-SanJuan 1998; Ibekwe-SanJuan & SanJuan, 2004), we propose enhancements to the linguistic component (adding of semantic relations) and explore a new application of our system, i.e., thesaurus construction and maintenance. The idea is to fine-tune the relations according to the target application. For thesaurus construction and maintenance, we select linguistic operations that induce explicit semantic relations. The rest of the paper is structured as follows: we first review previous works on automated methods for thesaurus construction (section §2); section three (§3) describes our methodology for extracting terms, identifying semantic relations and clustering them based on these relations; illustration of the method is carried out on a collection of scientific abstracts from the information retrieval field in section four (§4); section five (§5) shows how the proposed structure can assist thesaurus construction or maintenance.

2. Automatic or semi-automatic thesaurus construction: state-of-the-art

Two main tasks are involved in thesaurus construction : term collection (choice of domain concepts to be incorporated into the thesaurus) and classification (deciding on the hierarchy and assigning concepts to this hierarchy). Earlier works on semi-automatic or automatic thesaurus construction have had to address these two tasks. These works divide into two main approaches : statistical and linguistic approaches.

2.1 Statistical and clustering approaches

Unsurprisingly, answers for the term collection task came from the IR field (Salton &

[1] This first sentence is a quotation of Miller (1997). Thesaurus construction: problems and their roots, *Information Processing & Management*, vol. 33, n° 4, pp. 481-93.

McGill, 1983) where different versions of the IDF index (Inverse document frequency) have been formulated to select index terms from documents. Adapted to the thesaurus construction problem, the assumption is that frequently co-occurring words within a text window (sentence, paragraph, whole text,...) point to some semantic cohesiveness. A considerable amount of work in the computational linguistics field aiming to identify "collocations" are based on similar assumptions (Smadja, 1993). Following this tradition, Church and Hanks (1990) used a clustering technique to produce classes of words found in the vicinity of one another. However, the co-occurrence approach to unit extraction is not error free. Some frequently co-occurring units may be syntactically invalid. Much depends on the size of the window within which the collocations are sought. Once the "index" terms are extracted, the next stage is how to form "classes" of semantically-coherent terms which could be useful for thesaurus construction. A possible answer is to cluster frequently co-occurring terms in order to form classes but like the extraction of the units themselves, the statistical approach cannot make explicit the relations between the units in the same class nor can it consider word order. Often, the units extracted were lone words whereas terms, especially in technical domains are multiword units (MWU). Lin (2002) proposed a clustering method for identifying different senses of lone words. This approach is useful for identifying groups of synonymous words which share a generic/specific relation with the parent concept. Precision is measured by comparing a cluster (a word sense) to a WordNet[2] sense of the word. However, as most statistically oriented methods, his method needs very high word frequencies. As evidenced by previous authors (Lancaster, 1998 ; Schneider & Börland, 2004), methods based on statistical co-occurrence of text units are bound to produce semantically non-motivated classes which will require a lot of human intervention before they can be of use for thesaurus construction. For this reason, the results from statistical co-occurrence methods are called "*first-order word associations*" (Grefenstette, 1994). On the other hand, classes produced by statistical methods have been applied to information retrieval to enhance recall with a certain degree of success. Indeed, a user interested in a word may also be interested in those frequently appearing with it.

An alternative approach to thesaurus construction and maintenance through bibliometrics has been suggested by Rees-Potter (1989), followed up by Schneider & Börland (2004). Bibliometrics is the quantitative study of scientific publications. As such, it relies on several statistical measures (distance and similarity measures) and data plotting techniques (multidimensional scaling). The most widely used bibliometric methods are the co-citation (Small, 1999 ; White & McCain, 1989) and the co-word (Callon *et al.*, 1983) analyses. The basic hypothesis on which these methods are based is that frequently co-cited documents or frequently co-occurring keywords identify closely related information units (references, journals, authors). These methods employ different clustering techniques to map out clusters of co-cited documents or co-occurring terms onto a 2D space. The layout of the clusters is often interpreted in terms of core and peripheral research fronts. The co-cited documents are regarded as "concept symbols" to which the co-citing documents point. Usually, the results of bibliometric studies are used for science policy making to rank authors, laboratories or journals using tools such as the Citation Index. A lot of debate has risen over the use of co-citation studies for science evaluation policy. Many researchers have questioned the validity of the co-citation hypothesis. It is not the object of this paper to air these views but an interested reader can find a detailed account in Schneider & Borland (2004).

What we are interested in here is the claim that bibliometric methods can be applied to thesaurus construction and maintenance. This hypothesis is based on the assumption that since co-citation analysis links groups of related works, this can lead to grouping documents of related contents. Accordingly, Rees-Potter (1989) sought to identify conceptual changes in two domains, sociology and economy. She used co-citation and citation context analyses to identify candidate thesaurus terms from a corpus. Citation context analysis consists in examining the textual surroundings of a citation in order to extract terms which could shed some light on the context of

[2] WordNet (Fellbaum, 1998) is an electronic semantic database where lone words are organized into sets of senses called "synsets". A synset can be considered as a cluster of words depicting a particular sense of a word.

citation. However, as this was done manually, it was time consuming and was not followed up by any implementation. Schneider & Borland (2004) followed up this idea by using different tools to automate the stages involved in co-citation analysis. After performing a co-citation analysis on a corpus of texts in the field of periodontics, the authors semi-automatically performed citation-context analysis on a sample of citing documents in order to extract candidate noun phrases (NPs) around a "concept symbol" (a cited document). This enabled them to obtain a "*concept symbol word profile*", i.e. a list of frequently occurring words or phrases attached to a cited document. The authors then sought for conceptual changes over time by comparing these "concept symbol word profiles" over three time periods. Furthermore, the concept symbol word profiles were mapped onto a 2D space using clustering and visualization techniques (co-word analysis, multidimensional scaling and network analysis). The co-word analysis measures the strength of association of each pair of words or terms as the frequency of their co-occurrence divided by the frequencies of their separate occurrences in a given context (usually a document). This measure is normalized in several ways (Inclusion index, cosine, Jaccard index). Here the documents are replaced by the citation contexts from where the concept symbol profiles were extracted. Multi-dimensional scaling (MDS) is a data plotting technique which generates a visual display of similarity matrices. Like most data analysis techniques, MDS has the major problem of plotting data from *n* dimensional spaces onto a 2D or 3D space. This is usually done at the price of some distortion or information loss. Moreover such techniques cannot scale up to handle very large matrices. An algorithm based on MDS was used by Schneider & Borland (2004) to plot several maps of co-cited documents and corresponding co-words found in their citation contexts. The authors then looked for visual changes in these maps across different time periods. It was not clear from the study whether the candidate terms extracted actually shared any semantic relations as no examples of such terms were provided. Usually, the maps produced by bibliometric methods are topographic representations of research topics or actors in a field and cannot label explicitly the semantic relations between each concept pair. More empirical evidence will be needed in support of the claim that bibliometric methods can be used for thesaurus construction and maintenance.

2.2 Linguistic-oriented approach

We call the methods found here "linguistic-oriented" because the main foundation for selecting terms and relations are based on linguistic criteria. However, some of the methods make use of statistical measures to select terms or to form classes of terms. The linguistically-oriented approach to thesaurus construction can be broken down into three specific approaches :

i- functional similarity (distributional approach),
ii- internal syntactic evidence,
iii- corpus-based semantic relation markers and external semantic resources.

The first and second approaches are related and are often combined. The distributional approach stems from Harris (1968) who stipulated that the more similar the distributional context of a syntactic unit (subject, object,..), the more likely that the units are synonyms. A distributional context consists of the syntactic units and the grammatical function of these units surrounding the item under study. This assumption is basically the same as the statistical approach, except that functional or grammatical constraints are imposed on co-occurring text units. In contrast to the statistical co-occurrence approach, classes of terms produced through linguistic criteria can be qualified as "*second-order associations*" since they take into account not only word-order but also the grammatical functions of the component words of a phrase or its syntactic function around another phrase (like the verb phrase). Hindle (1990) and Grefenstette (1994) hypothesize that text units (mainly NPs) frequently occurring in similar syntactic functions (subject, object) form some sort of 'semantic classes'. Such functional similarity is bound to identify synonyms or some kind of "relatedness" rather than hierarchical relations. In his study, Grefenstette targeted the synonymy relation and used a similarity measure in order to relate contexts to each other. In such an approach,

the frequency threshold above which co-occurring functions are considered influences the results. Moreover, not all thesaural relations are addressed and the methods do not usually produce a hierarchy or a structured organization of domain terms.

The internal syntactic evidence approach relies on the grammatical function of component words within a term. Several studies have been done on organizing terms by "head-modifier" relations (Ruge, 1992; Woods, 1997; Grefenstette, 1997). The head is the noun focus and is the last noun in a compound structure (*thesaurus construction system*) or the noun preceding the preposition in a syntagmatic structure (*system for thesaurus construction*). The idea of the 'modifier-noun' function distinction is to form classes of nouns (heads) that share same modifier or alternatively classes of modifiers around the same head noun. The former would portray sets of related concepts that share similar properties (modifiers) while the second would portray sets of related properties around the same concepts (heads). Usually, a similarity coefficient is calculated using the mutual information index (Church & Hanks, 1990) in order to select the sets of heads (resp. modifiers) which will be put into the same class. In this way, Grefenstette (1997) produced classes for the word "*research*", depending on the grammatical function of the word: types of research (*market research, recent research, scientific research*) and research things (*research project, research program, research center*), (Sanderson & Croft, 1999). In the first case where research is the head noun, the three terms tell us something about "*research*" but without making explicit the exact semantic relation. As with the distributional approach, internal evidence applied in this way identifies loosely-related sets of terms which may share several semantic relations: "*research project*" and "*research program*" maybe considered as synonyms whereas a "related" relation holds between the former two and "*research center*" (place where *research project* and *programs* are carried out). The above two approaches are useful in information retrieval tasks such as query expansion but require more filtering before they can be used for thesaurus construction. Also, the output of these approaches is not a hierarchical structure but rather horizontal one (synonyms, related terms). They need to be completed by other linguistic cues like "subsumption" (lexical inclusion) in order to identify also hierarchical relations.

Aside from internal evidence gained from the term's internal structure itself, other means of acquiring explicit semantic relations from texts have been investigated. These means can be categorized either as endogeneous (corpus-based) or exogeneous (use of external semantic resource). The distributional and internal evidence approaches fall under endogeneous approach. A third type of endogeneous approach to semantic relations acquisition relies on a series of surface relational markers identified in each language. The underlying hypothesis is that semantic relations can be expressed via a variety of surface lexical and syntactic patterns. According to Condamines (2002: 144-145), this idea can be traced back to Lyons (1978) who used the term *formulae* to refer to terms linked by a hyperonymic relation, and also to Cruse (1986) who spoke of *diagnostic frames*. Hearst (1992) identified a list of hypernym/hyponym markers which have come to be widely used for corpus-based semantic relations acquisition. Examples are the sequence "*such NP1 as NP2, NP3 and /or NP4*" where NP is a noun phrase. Morin & Jacquemin (2004) adapted these hypernym/hyponym (generic / specific) markers to the French language and used them to acquire relations from corpora with the aim to assist thesaurus construction. However, for their method to function, it needs a bootstrap of manually defined pairs of semantically-related words. Then combining three separate tools: Prométhée (Morin, 1998), Acabit (Daille, 1996) and Fastr (Jacquemin, 2001), the method incrementally learns other relational contexts in the corpus. Prométhée uses a bootstrap of handcrafted lexico-syntactic patterns to acquire specific semantic relations between terms in a corpus. This tool extracts generic/specific term candidates incrementally. Acabit is a term extractor based on symbolic and statistical features. The symbolic features are used to select likely morpho-syntactic patterns of terms (N-N, N-of-N, ...) while statistical measures are used to sort the selected candidates by order of likelihood. Fastr is a term variant generator. Starting from a lemmatized list of terms, Fastr finds their morpho-syntactic and semantic variants in a corpus. An inherent limitation of the relational markers as defined by Hearst is that they are intra-sentential, i.e. the related terms have to occur in the same sentence. This leads

to the loss of semantically-related terms which do not occur in the same sentence. As a remedy to this limitation, Morin & Jacquemin (ibid.) perform an additional expansion of hypernym links by mapping related one-word terms onto multiword terms. For instance, given a link between "*fruit*" and "*apple*", similar links between the multi-word terms "*fruit juice*" and "*apple juice*" are extracted. This relationship is established on the basis that:

(i) the two terms share the same head (*juice*),
(ii) the substituted words have the same grammatical function (modifiers)
(iii) the substituted words be semantically close.

This last information is obtained from an external semantic resource, in this case the Agrovoc[3] thesaurus but it could also be learned from a general resource such as WordNet (Fellbaum, 1998). Using the hierarchy of single words from the Agrovoc thesaurus, the authors were able to project links between multi-word terms and thus build some partial hierarchies. However, they observed some erroneous links projected in this way. For instance, a link established between "*pêche*" (fishing or peach) and "*fruit*" was erroneously transferred to "*produits de la pêche*" (fishery products) and "*produits à partir de fruits*" (products from fruits) because of the ambiguity of "*pêche*" in French. According to the authors, the rate of error among the links generated in this way is rather low because this was done on very specialized domain where many of the multi-word terms were non polysemous. However, they admitted that this method cannot acquire all possible links between a single term and its hyponyms because multi-word terms are less frequent in a corpus. Their methodology is considered as a "context-based assistance to the extension of a thesaurus" (Morin & Jacquemin, 2004). Their study constitutes, to our knowledge, the most elaborate attempt to use linguistic knowledge instead of statistical one to build or update a thesaurus.

Our methodology shares some common features with Morin & Jacquemin's study. First is the fact that we also extract terms and their variants (morphological, syntactic and semantic) which constitute the "index units". We also establish several linguistic relations between these terms using internal evidence (the term structure itself) and contextual evidence (relational markers). Complementary external resources can be used where available to increase the number of semantic links acquired. Internal evidence can be viewed also as a way of projecting relations between terms that are not necessarily in the same sentence. Where our method differs from the work of these authors is in the way in which the terms are organized: we build classes of terms based on a clustering algorithm which not only groups synonymous terms, generic/specific terms but also associated terms. This is done iteratively such that specific semantic relations can be integrated at different iterations of the clustering algorithm. Thus these classes represent, for a given term, its closest neighbours in terms of semantic relations. A thesaurus builder can then decide how to place these terms and their links in the target hierarchy.

3. Our methodology : combining NLP and Clustering techniques for thesaurus construction

This method has resulted in the development of a term clustering system named TermWatch, tested on several corpora for various information-oriented tasks such as science and technology watch (Ibekwe-SanJuan & SanJuan, 2004), ontology population (SanJuan *et al.*, 2005). The system comprises three major components: a linguistic component which extracts terms and identifies relations between them, a clustering component that clusters terms based on the explicit linguistic relations identified, an integrated visualization interface which enables the user to explore the organization of clusters. We first present the text collection used in our study before describing the methodology.

3.1. Corpus constitution

Illustration of the methodology will be done on a collection of scientific titles and abstracts of papers published in 16 information retrieval journals between 1997-2003, called *IRCorpus*

[3] A multilingual thesaurus for indexing and searching agricultural databases managed by the FAO.

henceforth. *IRCorpus* comprises 455 000 words which were downloaded from the PASCAL bibliographic database maintained by the French institute for scientific and technical information (INIST/CNRS). The corpus was collected on the basis that they were summaries of research published by leading IR journals over eight years and as such should reflect important terminology in the IR field. The list of source journals is given in the appendix.

3.2. Linguistic component

This component relies on shallow NLP to extract terms and relates them through various relational devices. Terms are linguistic units (words or phrases) which taken out of their contexts, refer to existing concepts or objects in a given field. In other words, terms are choice linguistic units, rich in information content because they are used by experts in a field to name the objects or concepts of that particular field. A lot of research has been done on automatic term extraction by computational terminologists (see Jacquemin & Bourigault, 2003 for a review). However, focus has now shifted towards automatic terminology structuring for various applications (ontology learning and population mostly) as shown by the recent issue of the journal '*Terminology*' (Ibekwe-SanJuan, Condamines, Cabré, 2005).

3.2.1 Term extraction

It is an accepted fact that most terms appear as noun phrases (NPs) although some verbs and prepositional phrases can be terms. We currently extract only terminological NPs which are multiword units (*information retrieval system ; peer review of technology-experts*). Term extraction is performed using the LTPOS tagger and LTChunker developed by the University of Edinburgh. LTPOS is a probabilistic part-of-speech tagger based on Hidden Markov Models. It has been trained on a large corpus and achieves an acceptable performance. Since LTChunker only identifies simplex NPs without prepositional attachments, we wrote contextual rules to identify complex terminological NPs. The term extraction process is exhaustive in that frequency is not used as a filter. Only badly formed candidates from the morpho-syntactic basis are dropped from further processing. This ensures that we will be working on all the candidate terms from the corpus and not on a subset as is usually the case with approaches based on statistical measures. Selection will be done through linguistic means. Terms which do not share any semantic relations with others cannot be further processed. Likewise, terms which are too loosely bound to the terminological network may disappear from the final clusters obtained. Term extraction on the *IRcorpus* yielded 44 665 candidate terms.

3.2.2. Acquiring semantic relations between terms

The issue of establishing relations between text units is not a recent one. In the computational terminology community, a paradigm has emerged which is known as the “variation paradigm”. As more and more corpus-based evidence become available, terms are no longer considered as “fixed” units but as dynamic units which vary under several linguistic influences. Variation denotes the fact that a term may appear under different forms and that locating these different forms is essential for many information-oriented tasks like query expansion, question-answering, information retrieval, indexing, terminology acquisition and structuring, ontology population. Variations are not rare phenomena, they affect about 35% of terms in a domain. Variations can take place at different linguistic levels: morphological, syntactic or semantic, thus making their identification impossible without integrating NLP techniques. Capturing these variations enhances the representation of the actual state of a domain's terminology. Also identifying and structuring terms along variation lines enhances the understanding of the conceptual relations between a domain's concept and hence represents an initial step towards knowledge organization.

Term structuring through variations has been explored for various applications like building domain lexical resources from corpora (Daille, 2003), automatic thesaurus building (Morin & Jacquemin, 2004), information retrieval, question-answering (Dowdall *et al.,* 2003), science and technology watch (Ibekwe-SanJuan & SanJuan, 2004), the list not being closed.

Here we combine several linguistic approaches to semantic relations acquisition:

i) internal evidence (morpho-syntactic variations),
ii) contextual evidence (lexico-syntactic patterns),
iii) external resource (domain thesaurus, ontology or general purpose semantic resource).

(i) and (ii) stem from endogeneous corpus-based approach while (iii) is an exogeneous approach to semantic relations acquisition. We thus propose a comprehensive approach in order to complement the shortcomings of each approach and increase the number of relations mined between corpus terms. This constitutes an enhancement to our previous studies where only the first approach was used (internal evidence).

- *Internal evidence or morpho-syntactic variations*

This relies on the structure of the terms themselves to establish relations between them. The relations identified here rely on two morpho-syntactic operations: expansions and substitutions. Expansions denote the fact that term1 is subsumed in term2 either by the addition of modifier or head words. For instance, between "*academic library*" and "*hellenic academic library*", the addition of modifier words (regardless of their number) tends to create a generic/specific relation (the longer variant is more specific) while in "*british library*" and "*british library restoration*", the addition of a new head word leads to a horizontal relation thus creating loosely "related terms". If a chain of modifier expansions is found in the corpus, this can lead to building a hierarchy from the most generic term to the most specific.

Substitution operates on terms of equal length in which one and only one element is substituted, either in a modifier or in a head position. This idea was exploited in Grefenstette (1997). Substitutions engender horizontal relations between terms. Therefore, the resulting conceptual relation is a more general "relatedness". Modifier substitutions can denote members of the same concept family with alternative qualifications, siblings in an IS_A hierarchy. The conceptual shift engendered by head substitutions, on the other hand, links different IS_A hierarchies at the same level of specificity.

The semantic links acquired through expansions and substitutions are effected through pure lexical association, i.e., the fact that two terms share some common elements. This enables us to capture quite a considerable number of links between corpus terms. The following figure shows the type of hierarchies that can be obtained through internal evidence.

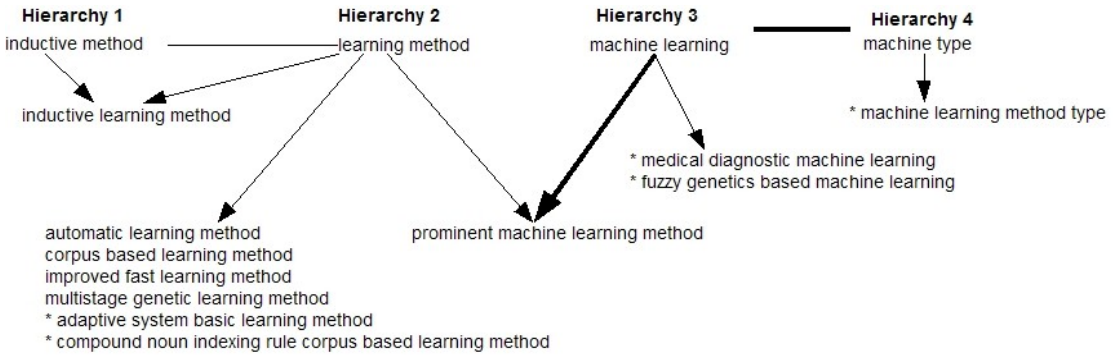

Figure 1. Example of semantic links generated by lexical association.

Simple lines denote modifier substitutions while simple arrows denote specific links generated by modifier expansions. Bold arrows denote head expansion links while bold lines denote head substitutions. The terms preceded by a star may not be direct hyponyms of the generic term since they contain many modifier words. There may be intermediary terms between them and the topmost generic term but which were absent in the corpus.

Lexical association enables us to obtain a polyhierarchical structure where a term may belong to two hierarchies as is the case for "*inductive learning method*" and "*prominent machine learning method*". While the association link between "*inductive method*" and "*learning method*" seems

understandable (both can be attached to a hierarchy of "*methods*" thus making them co-hyponyms), the link between "*machine learning*" and "*machine type*" (head substitution) is less obvious and less interesting. Given their definition, substitutions are the most prolific and create sets of loosely-related terms whose significance is not always clear for the domain specialist. For this reason, they will be subjected to further filtering before clustering in order to retain only semantically-motivated substitution variants (§2.2.3). 26 128 terms were involved in relations through internal evidence, thus more than 58% of the total number of extracted terms (44 665).

- *Contextual evidence or lexico-syntactic relational markers*

While internal evidence undeniably enables the acquisition of a certain number of semantic links such as generic/specific, the approach is inherently limited in that it cannot capture conceptually related terms which do not share any lexical element. For instance, *AltaVista* and *search engine* will not be related, nor will *car* and *vehicule* be, whereas both pairs obviously share a generic/specific relation. In other words, internal evidence cannot detect semantic variants when the relation is materialized by other linguistic devices aside from lexical association (common words between the two terms). To ensure a maximum capture of semantic relations between domain terms, we need a complementary approach which relies on contextual relational markers. These markers have been studied by Hearst (1998) among others. The generic lexico-syntactic patterns amongst which generic/specific relations can occur are represented by the following regular expressions :

H1 *: such* NP_0 *as* NP_1 (<cc>)+ NP_2 …, (<cc>)* NP_n

H2 : NP_0 *such as* NP_1 (<cc>)+ NP_2 …, (<cc>)* NP_n

H3 : NP_0 (<cc> | NP_1) *like* (NP_2 | <cc>)+ NP_n

H4 : NP_0 (<cc> | NP_1) (*particularly* | *especially* | *in particular*) (NP_2 | <cc>)* NP_n

H5 : NP_0 (<cc> | NP_1) *including* (<cc> | (NP_2 | <cc>*)) NP_n

H6 : NP_0 ([**;**| **:**]) NP_1 [parenthesis] (**,**) NP_2* (<cc>) NP_n

H7 : NP_0 (<cc> | NP_1) [,] (<&>) *other* NP_2 …, (<cc>)* NP_n

H8 : NP1 (**,** | namely) NP_2 …, (<cc>)* NP_n

where NP can either be a simplex or a complex noun phrase with a prepositional (PP) attachment ; "<cc>" is a coordinating conjunction element such as "*and, or, comma*" ; "*" is Kleene's star ; "+" means that there must be at least one occurrence of the preceding element. Brackets encase optional syntactic structures (NP) or lexical elements (cc). The words in brackets are optional. The lexical elements in angle brackets are not terminal categories. When such relational markers are found in the text, the surrounding NPs are marked and extracted with the labeled semantic relation. For instance, pattern H2 would apply to the sentence below.

(1) challenging requirements in [HYPER] non-traditional applications <u>such as</u> [HYPO] geographic information systems ([HYPO] GISs), [HYPO] computer-aided design ([HYPO] CAD), and [HYPO] multimedia databases.

This yields five pairs of terms in a generic/specific relation. The specific terms are in turn related as co-hyponyms. Graphically, we obtain figure 2 where arrows indicate the hierarchy engendered by the generic/specific relations. Symmetrical relations denoting co-hyponyms are shown by simple straight lines. Furthermore, since "*GIS*" is an acronym of "*geographic information system*" and "*CAD*" an acronym of "*computer-aided design*", the two pairs are semantic equivalents. This is represented by bold lines.

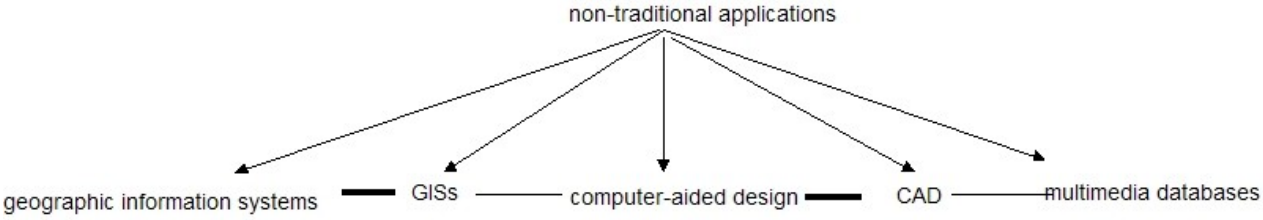

Figure 2. Hierarchical and horizontal links obtained by a lexico-syntactic pattern.

Certain lexico-syntactic patterns signal the presence of synonyms relations between terms. Authors writing up a text make use of different naming mechanisms in order to introduce alternative terms for the same concept. Amongst the most cited patterns found in the literature (Suarez & Cabré, 2002), we tested the following :

S1 : NP_1 (also | equally | now) *called* (now, also equally) NP_2

S2 : NP_1 (also | equally | now) *known as* NP_2

S3 : NP_1 (also | equally | now) *named* NP_2

S4 : NP_1 (also | equally | now) *viewed as* NP_2

S5 : NP_1 (also | equally | now) *termed as* NP_2

S6 : NP_1 *termed* (for | as) NP_2

S7 : NP_1 *referred to as* NP_2 (<cc> | NP_3)

S8 : NP_1 (<have> (also | equally | now) |<be>) defined as NP_2

S9 : NP_1 <be> no other than NP_2

The infinitive verbs appear in inflected mode (tense, number). We observed that these patterns are subject to variations, for instance, the presence of an optional punctuation mark (the comma) appearing before the cue words (*called, named*, ...). Also the cue words can appear in parenthesis with one of the synonym terms. It is interesting to note that while the adjective form "*named*" functions as a synonymy marker, its adverbial form (*namely*) is used in a generic/specific relational pattern (H8). Given the following sentence, pattern S2 will mark and extract the terms "*mathematical operation*" and "*convolution*" as synonyms.

(2) This combination is performed by a [SYN] mathematical operation ***known as*** *[SYN] convolution*.

We manually validated the contexts extracted by the two types of relation markers and measured their precision and recall. Generic/specific markers enabled us to extract 571 contexts relating 1162 pairs of terms. A context with all candidate pairs was accepted if the relation was judged semantically sound, if not both the context and the candidate terms were rejected. 77% out of the 571 contexts contained relevant relations while 23% were discarded.

Synonymy markers enabled us to extract 107 contexts containing 146 pairs of synonym terms. As synonymy is often a binary construction (*term A is also called term B*), it hardly involves enumeration devices unlike the generic relation markers, hence the number of synonym pairs is close to the number of contexts. Out of the 107 contexts extracted by synonym markers, 92% represented valid synonymy relations while 8% were discarded.

Recall consisted in checking if any of the relation contexts was missed by the lexico-syntactic patterns. Given the size of the corpus (455 000 words), we could only carry out this task on a small portion. The first 200 out of the 3355 abstracts were manually checked for other generic/ specific or synonym relations which were not extracted by our lexico-syntactic patterns. The result was compared against the list of generic/specific terms acquired by the patterns on this portion of the corpus.

The rate of recall was found to be high (86%) for generic/specific relations whereas it was slightly lower for synonym patterns (77%).

- *External semantic resource*

This third approach to semantic relations acquisition completes the other two (internal evidence and contextual relation markers) by enabling the acquisition of relations which were not signaled by endogeneous markers. This can be done via a domain thesaurus, ontology or a general language resource such as WordNet (Fellbaum, 1998). WordNet is a general language semantic database with three categorial hierarchies: noun, verb and adjective. WordNet contains generic/specific, synonym, meronym and "relatedness" as well as other semantic relations. Words are organized in "synsets" in WordNet, a synset being a class of words representing the same sense. Typically, a polysemous word may have up to six synsets. Using an external resource is also a means of filtering the substitution relations acquired by internal evidence. In the *IRcorpus*, substitutions were filtered in order to obtain two subtypes: strong and weak substitutions. Strong substitutions are those whose substituted words belonged to the same synset. Weak substitutions are all the other lexical substitutions involving terms of length ≥3. The idea was to filter out binary substitutions as these were very prolific and create a lot of noise in later processes. Some examples of strong an weak substitutions are given below.

Substitutions	Nb. terms	Nb. links	Examples
Strong_Sub (WordNet filtered)	1184	1849	subject_JJ categorization_NN subject_JJ classification_NN
			recall_NN measure_NN recall_NN measurement_NN
			retrieval_NN outcome_NN retrieval_NN result_NN
			computer_NN graphics_NN domain_NN computer_NN graphics_NN field_NN
			*multiple_JJ language_NN *multiple_JJ way_NN
Weak_Sub (length ≥3)	5855	21 684	citation-based_NN retrieval_NN system_NN document_NN retrieval_NN system_NN face_NN retrieval_NN system_NN image_NN retrieval_NN system_NN information_NN retrieval_NN system_NN online_NN retrieval_NN system_NN text_NN retrieval_NN system_NN

Table 1. Strong and weak substitution variants.

We were able to filter out the majority of the substitutions acquired through internal evidence while highly improving the precision. The starred examples illustrate the gap between what can be accepted as general language synonyms (*way* / *language*) and what can be considered as synonyms in a specialized domain when using a general vocabulary external resource such as WordNet. Here, *multiple language* and *multiple way* are surely not IR terms and not synonymous in this field.

The combination of different linguistic approaches described in this section, enabled us to acquire automatically several thousands of semantic links between corpus terms, each approach complementing the limitations of the other. This constitutes a comprehensive method for domain knowledge acquisition and structuring. Figure 3 below summarizes the process of semantic relations acquisition involved in the linguistic component.

Currently, two means of acquiring semantic relations have been implemented in the TermWatch system: internal evidence and external evidence using WordNet. Acquiring semantic relations through relational markers (contextual evidence) is yet to be implemented.

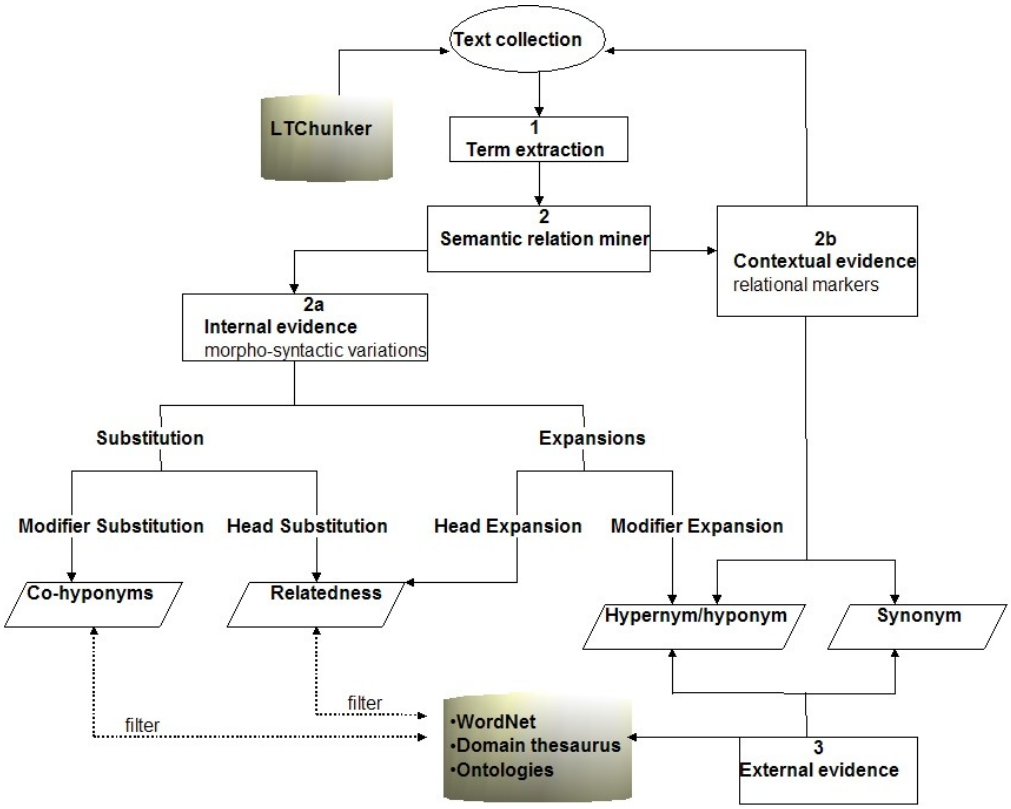

Figure 3. The linguistic component.

3.3. Clustering component

Clustering is a data analysis technique used to generate automatically groups of similar objects. Objects in a cluster should have a greater degree of proximity than those in different clusters. Similarity is usually defined as a co-occurrence function in a text window (sentence, paragraph, whole document). There exists in the literature different ways of classifying data analysis techniques. A distinct family of methods is factor analysis comprising principal component analysis (PCA) and multi-dimensional scaling (MDS). However, as we have pointed out earlier, factor analysis involves some data loss and distortion during the plotting. Also, it does not scale well to very large and sparse matrices. Another family of data analysis methods is clustering. There are several clustering algorithms which divide into two families: hierarchical or partitioning algorithms. Hierarchical clustering algorithms are either agglomerative (they start from atomic units and merge them iteratively into nested clusters) or are divisive (they start by putting all objects into one cluster and then progressively divide it into several clusters). Partitioning algorithms (k-means) on the other hand have to find a partition where each object is affected to the cluster with the closest "centroid" or center. Partitioning algorithms and factor analysis both rely on distance measures. As such, the analyzed units have to lend themselves to the definition of a distance like chi-2 or the Euclidian distance. Hierarchical algorithms on the other hand use different mechanisms to interpret the similarity measures between objects in a matrix: single link, complete link or average link. In single link clustering (SLC), an object need only have one link for it be integrated into a cluster. This leads to the well-known and undesirable "chain effect", i.e., very long clusters whose internal links may be weak in some places. By opposition, complete link clustering imposes several links between an object and members of a cluster. This leads to more cohesive clusters but also to forming many singletons (clusters of lone items).

Graph theoretic algorithms are better suited to the nature of our sparse data. The graph theoretical algorithm that has the strongest mathematical properties is clearly SLC because it entails forming connected components from the graph of similarities. However, SLC has the major

drawback of the chain effect, thus making it only adapted for special datasets where the desired clusters have this property. Our clustering technique follows the SLC principle but avoids the chain effect by interpreting similarity values in a relative manner, i.e., two vertices are merged if the similarity values are higher than the ones surrounding the vertices. Thus at a given iteration, several sets of vertices can be merged at different similarity values. Similarity values are not considered as an ordered set. In other words, we consider the local maximums of the similarity function whereas the classical SLC considers similarity as an ordered set and at a given iteration, will cluster units only above a given threshold. Thus at a given iteration, classical SLC will merge only sets of items with the same similarity value. Some implementations of SLC limit the chain effect by imposing a maximum size on clusters (Callon *et al.*, 1983). However this comes at a cost: the output of the clustering is no longer unique. A change of cluster size or of the order in which units are clustered affects the results.

Because we cannot define a distance on our datasets (terms) and the semantic relations they share (there is no way to define triangular inequality), because frequency information is not used in our clustering scheme, we required an algorithm which performs well on very sparse data and can differentiate between several symbolic relations. This excluded algorithms based on distance measures (K-means, factor analysis). Similarity in our method is not measured on a co-occurrence matrix. Rather clustering is performed on a graph of semantically-related terms, i.e., the semantic relations identified in the preceding stage. This is a radically different approach to current data analysis techniques. We describe below the functioning of the method.

3.3.1 Choosing the relations for clustering

The semantic relations identified by the linguistic component can be seen as dimensions from which interaction between domain concepts can be viewed. Depending on the application targeted, the user can choose the relations to be used for clustering and assign a role to them during the clustering process (§3.3.2 below). For instance, for domain terminology structuring where the focus is on obtaining semantically-tight clusters, it will be more relevant to give priority to filtered substitutions and modifier expansions within the variants acquired through internal evidence, together with the semantic relations acquired through contextual and external evidence. In other words, applications like thesaurus and ontology building need a careful selection of only those variants which are semantically close. Contrarily, when targeting applications like science and technology watch or query expansion, the focus is on "recall" or associations rather than on semantic precision. Hence, the tendency will be to include less semantically-motivated relations and form loose semantic clusters whose contents are suggestive. For science and technology watch, this could trigger off more explorations to understand the nature of the link by returning to source documents (Ibekwe-SanJuan & SanJuan, 2004). For query expansion, expanding cluster contents to related terms could be a way of suggesting other query terms to the user. The choice and role of each relation is done in an interactive interface. In between iterations, the user can change the role of any relation to better suit his/her purpose. We used the relations acquired by internal evidence and assigned them a role according to the semantic proximity induced between terms. Modifier expansions are likely to indicate generic/specific links, WordNet-filtered substitutions are likely to denote synonyms while head expansions point to "related terms". We thus selected strong substitutions (strong-Sub) and left expansion (left-Exp) as priority relations for gathering terms into connected components. We call these priority relations "COMP" relations. The remaining ones: weak substitutions (weak-Sub), right expansion (right-Exp) and insertions (ins) are second level relations called "CLAS relations".

3.3.2 Clustering semantic variants

The variation relations used are represented as a graph. We recall briefly the functioning of the algorithm. CPCL (Classification by Preferential Clustered Link) is a hierarchical two-step

extractor of clusters from a graph of term variants (Ibekwe-SanJuan, 1998).
One notable attribute of this algorithm is that the clustering begins not at the atomic level (term level) but at the component level. Components are obtained by grouping terms sharing COMP relations. A dissimilarity function *d* is calculated for every pair of components as the proportion of CLAS relations between them. More formally, *d* is an application in [0,1] defined for every pair (*i, j*) of components as follows :

i) $d(i,j) = 1$ if for every r in $\{1,\ldots,k\}$, $N_r(i,j) = 0$ and $d(i,j) = 0$ if i = j ;

ii) $d(i,j) = 1 / \sum_{r=1}^{k} \frac{N_r(i,j)}{|R_r|}$ where $R_1 \ldots R_k$ are CLAS relations and $N_r(i,j)$ is the number of links in R_r between *i* and *j*.

Once we have defined such a dissimilarity index on a graph, there exists a large variety of data analysis methods to cluster the set of vertices, but only few of them are computationally efficient for large sparse graphs. Sparse graphs have few edges compared to the number of vertices and consequently are difficult to cluster using usual hierarchical and *k*-means procedures from statistical clustering algorithms. The clustering stage consists in merging iteratively components that share many CLAS relations. The user can set the number of iterations and the minimal dissimilarity index to be considered or let the algorithm converge and then choose the results of a given iteration.
The results obtained on the *IRcorpus* is taken at the first iteration, no threshold for minimal dissimilarity index was set. The system built 674 clusters with 1595 components and a total of 5632 terms. 15 clusters had more than 30 terms and 517 clusters had less than 10 terms. The following section analyzes the structure obtained.

4. Mapping the information retrieval corpus

We first give a global view of domain themes as mapped by TermWatch. The clusters are graphically displayed using *Aisee*, an integrated visualization tool. We will comment on the layout and content of specific classes and show how this can assist thesaurus construction and maintenance.

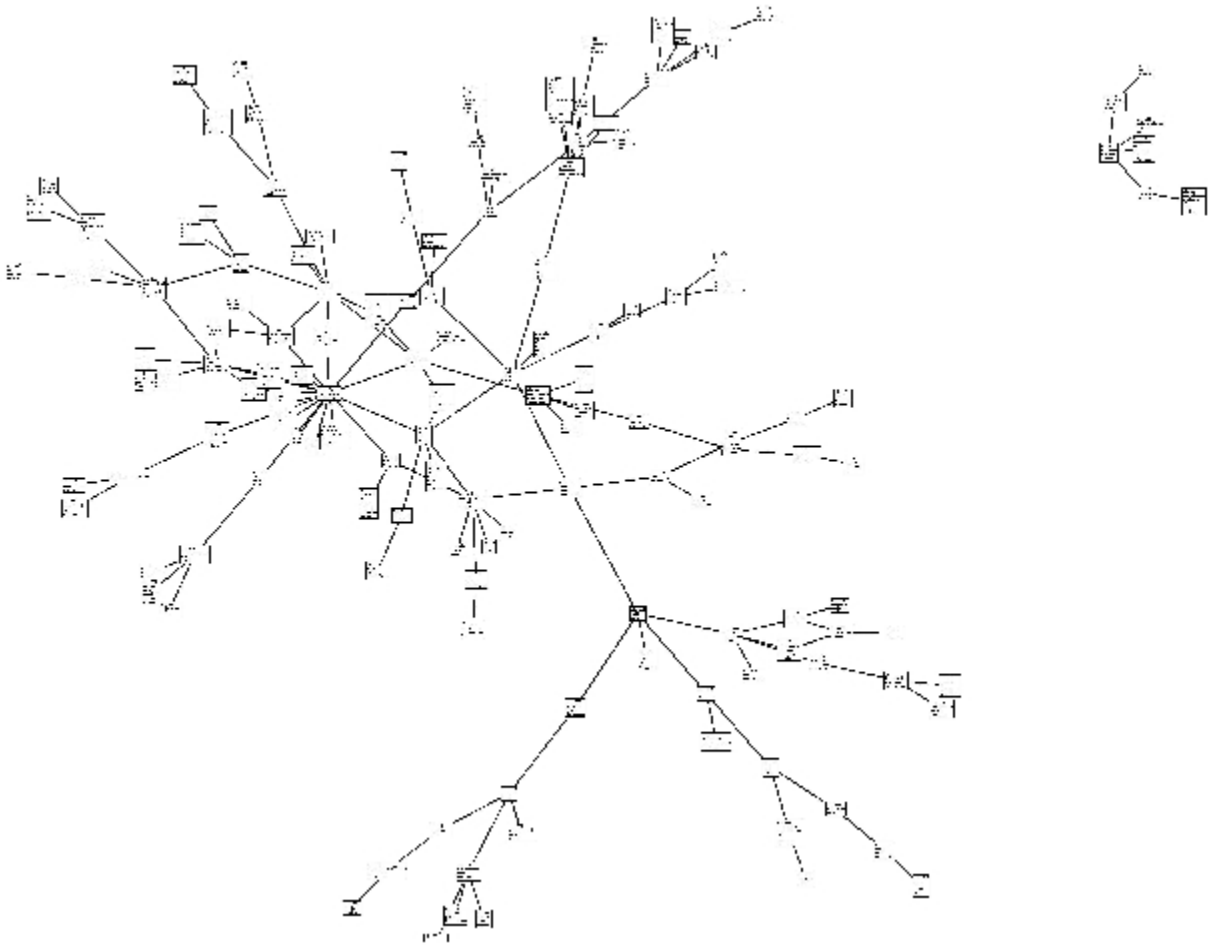

Figure 4. Global view of the map of domain themes.

The global image is too small to be legible. The user can use zoom functions to explore different areas of the graph. The core position is occupied by a cluster labeled "*information retrieval*". While this may not be surprising, it is not altogether a trivial finding because the corpus was not built using keywords but using journal names while the clusters were built from the summaries and titles of published papers. So it cannot have been foreseen that researchers actually employed the term "*information retrieval*" to describe their research. This cluster was also the biggest with 135 terms, among which 84 were in the same component labeled "*Information retrieval system*".

4.1 Exploiting cluster content for thesaurus construction or maintenance

We now illustrate how the semantic variants in clusters can assist domain knowledge organization tasks like thesaurus building and maintenance. An example of the organization that can be obtained from a cluster's content is given for cluster 141, labeled "*system transaction logs*". This cluster contains thirteen terms. Figure 5 below shows all the terms in this cluster, structured according to the semantic relation induced by each linguistic operation studied.

Cluster 141. System transaction logs

- Log analysis
  - NT transaction log analysis
    - NT longitudinal transaction log analysis
  - RT transaction log
- Transaction log
  - NT Excite transaction log RT

NT Inquirus search transaction logs
NT month transaction log
TS month transaction log of SuperJournal
NT on-line transaction logs
NT search transaction logs
NT system transaction logs
NT transaction logs of search strategies
RT Transaction log analysis

Note.
Terms that are specific of another term, at the same hierarchical level are also associated terms, i.e.,
on-line transaction log, search transaction log, system transaction log, month transaction log, excite transaction log, etc are also related terms

Figure 5. Structuring the content of cluster "*System transaction log*" for thesaurus construction.

Currently, all the relations are induced by internal evidence (lexical association) and WordNet-filtered substitutions. They only require a shallow NLP processing. The proposed structure in Figure 5 is not built automatically yet, but this can easily be done by formatting cluster contents to suit thesaurus presentation. It suffices to rebuild the hierarchy of relations basing on the formal properties of the relations used for clustering. Modifier relations involving terms of different length imply anti-symmetrical links. This can be used to obtain the hierarchy of broader and narrower terms. Head relations involving terms of different length as well as substitution relations imply association (horizontal links). WordNet substitutions imply synonymy, thus "Used for/ Use" relations. However, the visual display of a thesaural structure is not a difficult issue as this will depend very much on the type of tool used by the thesaurus builder.

Conclusion

We have reviewed existing approaches to automatic thesaurus construction and maintenance. Our finding is that statistical approaches produce possibly related terms but without labeling the explicit semantic relation between them and without proposing a hierarchy of these terms. Linguistic-based approaches on the other hand bring solutions to the first problem but do not necessarily produce the required term hierarchies. Although tools for semi-automatic thesaurus and ontology population exist, they mostly offer assistance in producing sets of related terms, some producing partial hierarchies (Morin & Jacquemin, 2004). The task of actually placing the terms and the relations at specific nodes in the thesaurus remains manual. In line with these semi-automatic and linguistically-based approaches, we have proposed a comprehensive methodology which in addition to semantic relation mining, employs clustering to produce classes of semantically-related terms. This has the added advantage of offering a conceptual organization of the important domain terms contained in the corpus. This is a more meaningful term organization device than frequency or co-occurrence criteria. From empirical evidence gathered across several corpora, the clusters produced by TermWatch capture most of the relations necessary to build or update a thesaurus. A recent experiment reported in SanJuan *et al.* (2005) showed that the clusters can reflect, to some extent, a human semantic organization of domain concepts in the genomics domain. In this experiment, TermWatch's clusters were compared against a manually built ontology (GENIA ontology[i]). It was found that the biggest clusters had more than 40% of their terms from the same semantic category in the GENIA ontology. We still have to implement the use of contextual evidence and specialized domain resources to acquire semantic relations. This will increase further the number of relations mined in the clusters.

However, all the terms and relations found in the clusters may not ultimately be used for thesaurus building. Much depends on the level of specialization of the thesaurus, on the pragmatic choices made by the domain specialist, i.e., the level of genericity/specialization required in the thesaurus.

These constraints, which are external to the methodology, will weigh on the measure of usefulness of our semantic clustering approach for thesaurus construction and maintenance.
In this perspective, TermWatch's assistance lies mainly in automatically acquiring and organizing domain terms from text collections. The global thematic maps proposed will enable the thesaurus builder determine the important concepts and their hierarchy in the field (thesaurus root nodes). This is not a functionality normally supported by thesaurus building systems. Many studies have focused on tools for thesaurus building and maintenance. Ganzmann (1990) established a "check list" of system requirements both in terms of material (computer) and thesaurus functionality. For the latter, a dedicated interface is necessary for specifying the thesaurus global hierarchy. At the term level, some expected requirements were the enabling of enable term entry, term definitions, scope notes, term origin, examples of usage, facet grouping, language information in case of multilingual thesauri. Other required features concern consistency checks and various visual presentations of the thesaurus (alphabetic, hierarchical, KWIC, KWOC, facet grouping). Some organizations such as the American Association of College & Research Libraries also edit guidelines for thesaurus maintenance[ii] which conform to the corresponding ISO recommendations. Our system's output can be interfaced with an existing thesaurus building tools which already possess these functionalities. TermWatch will then feed the system with terms and relations which the user can accept or reject.

Acknowledgements
This research benefited from collaboration with Eric SanJuan, Lecturer in Computer sciences at the University of Metz (France) who implemented the TermWatch system.

References

Aitchison, J., Gilchrist, A., Bawden, D. (2000), *Thesaurus construction and use: A practical manual*, 4th ed., Aslib, London, 240p.

Condamines, A. (2002), "Corpus Analysis and Conceptual Relation Patterns", *Terminology,* Vol. 8 No. 1, pp.141-162.

Callon, M., Courtial, J-P., Turner, W., Bauin, S. (1983), "From translation to network : The co-word analysis", *Scientometrics*, Vol.5 No. 1.

Church, K.W., Hanks P. (1990), "Word association norms, mutual information and lexicography", *Computational Linguistics*, Vol.16 No. 1, pp. 22-29.

Cruse, D.A. (1986). *Lexical Semantics*, Cambridge: Cambridge University Press.

Daille, B. (1996). "Study and implementation of combined techniques for automatic extraction of terminology", in P. Resnik and J. Klavans (eds.). *The Balancing Act : Combining Symbolic and Statistical Approaches to Language,* Cambridge: MIT Press, pp. 49-66.

Daille, B. (2003), "Conceptual structuring through term variations", *Proceedings of the ACL-2003, Workshop on MultiWord Expressions: Analysis, Acquisition and Treatment,* Saporro, Japan, pp. 9-16.

Dowdall, J., F. Rinaldi, F. Ibekwe-SanJuan and E. SanJuan. (2003), "Complex structuring of term variants for question answering", in Bond, F. A. Korhonen, D. MacCarthy and A. Villacicencio (eds.). *Proceedings ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*. Sapporo, Japan, pp. 1-8.

Fellbaum, C. (1998), *Wordnet. An Electronic Lexical Database*. Cambridge, London: The MIT Press.

Ganzmann J.(1990), Criteria for the evaluation of thesaurus software, *International Classification,* Vol. 17*,* No 3/4, 148-157.

Grefenstette, G. (1997), SQLET:Short Query Linguistic Expansion Techniques, Palliating One-Word Queries by Providing Intermediate Structure to Text, *Proceedings of "Recherche d'Information assistée par ordinateur" (RIAO)*, pp. 500-9.

Grefenstette, G. (1994), *Exploration in Automatic Thesaurus Discovery*, Boston, MA: Kluwer Academic Publisher.

Harris, Z. S. (1968), *Mathematical Structures of Language*, New York: Wiley.

Hearst, M.A. (1992), "Automatic acquisition of hyponyms from large text corpora", *Proceedings of the COLING'92*, Nantes, pp. 539-545.

Hindle, D. (1990), "Noun classification from predicate argument structures", *Proceedings of the 28th Annual Meeting of the Association for Computational Lunguistics*, Pittsburg, PA.

Ibekwe-SanJuan, F., Condamines, A., Cabré, T. (eds.) (2005), "Application-driven Terminology engineering", *Special issue of Terminology : International journal of theoretical and applied issues in specialized communication*, John Benjamins, Vol. 11 No. 1, 200.

Ibekwe-SanJuan, F. and SanJuan, E. (2004), "Mining textual data through term variant clustering: the termwatch system", *Proceedings "Recherche d'Information assistée par ordinateur" (RIAO),* Avignon, pp. 487-503.

Ibekwe-SanJuan F., SanJuan E. (2002) From term variants to research topics. *Journal of Knowledge Organization (ISKO), Special issue on Human Language Technology,* Vol. 29 No 3/4, 181-197.

Ibekwe-SanJuan, F. (1998), "A linguistic and mathematical method for mapping thematic trends from texts", *Proceedings of the 13th European Conference on Artificial Intelligence (ECAI'98)*, Brighton UK, 23-28 August 1998, pp. 170-174.

Jacquemin, C., and Bourigault, D. (2003), "Term Extraction and Automatic Indexing", in R. Mitkov, (eds), *Handbook of Computational Linguistics*, Oxford University Press, pp. 599-615.

Jacquemin, C. (2001), *Spotting and discovering terms through Natural Language Processing*, MIT Press, 378p.

Pantel P., and Lin, D. (2002), "Discovering word senses from texts", *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2002,* Edmonton, Canada pp. 613-619.

Lyons, J. (1978), *Éléments de sémantique*. Paris: Larousse Universités.

Morin, E., Jacquemin, C. (2004), "Automatic acquisition and expansion of hypernym links", *Computer and the humanities,* Vol. 38, No. 4, pp. 363-396.

Morin, E. (1998), "Prométhée : un outil d'aide à l'acquisition de relations sémantiques entre termes", *Proceedings Traitement automatique des langues naturelles,* Paris, France, pp. 172-181.

Pedersen, T., Patwardhan, S., Michelizzi, J. (2004), WordNet::Similarity : Measuring the Relatedness of Concepts, *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04),* San Jose, CA, July 25-29, 4p.

Rees-Potter, L.K. (1989), Dynamic thesaural systems: a bibliometric study of terminological and conceptual change in sociology and economics with the application to the design of dynamic thesaural systems, *Information Processing & Management*, Vol. 25 No. 6, 677-91.

Ruge, G. (1992). Experiments on linguistically-based term associations, *Information Processing & Management*, Vol. 28 No. 3, pp. 317-32.

Sander, G. (1996), "Visualisierungstechniken für den Compilerbau", Dissertation, Pirrot Verlag & Druck.

Sanderson, M, Croft, W.B. (1999), "Deriving concept hierarchies from text", *Proceedings of the 22nd Annual ACM SIGIR Conference on research in Information Retrieval, Berkeley*, CA, 15-19 August, pp. 206-213.

Salton, G., McGill, M.J. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, NY.

SanJuan, E., Dowdall, J., Ibekwe-SanJuan, F., Rinaldi, F. (2005), "A symbolic approach to automatic multiword term structuring", *Computer Speech and Language* (CSL), *Special issue on Multiword Expressions*, Elsevier, 20p. [Forthcoming].

Schneider, J.W, Borlund, P. (2004), "Introduction to bibliometrics for construction and maintenance of thesauri", *Journal of Documentation*, Vol. 60 No. 5, pp. 524-549.

Smadja, F. (1993), "Retrieving collocations from text : Xtract", *Computational Linguistics* 19 (1), 143-177.

Small, H. (1999), "Visualizing science by citation mapping", *Journal of the American society for Information Science*, Vol. 50 No. 9, pp. 799-813.

Suárez, M., Cabré M.T. (2002), "Terminological variation in specialized texts: linguistic traces for automatic retrieval", *Proceedings VIII IberoAmerican symposium on Terminology*, October 28-31, 10p

White, H.D., Mccain K.W. (1989), "Bibliometrics", in M.E. Williams (ed.), *Annual Review of Information Science and Technology*, New York : Elsevier Science Publishers, pp. 119-186.

Woods, W.A. (1997), "Conceptual indexing: a better way to organize knowledge", Sun Labs Technical Report: TR-97-61, Sun Microsystems Laboratories, Mountain View, CA.

Appendix : The 16 IR journals used for the corpus constitution.

Column one is the journal rank, column two gives the number of bibliographic records per journal,

column three the proportion in the entire corpus and column four, the cumulative % and the last column, the journal name.

1	831	25%	831	25%	Information sciences
2	688	21%	1519	45%	J. of the Am. Soc. for Information Science and Technology
3	283	8%	1802	54%	Information processing & management
4	272	8%	2074	62%	Journal of information science
5	267	8%	2341	70%	Information systems management
6	175	5%	2516	75%	Journal of Documentation
7	176	5%	2692	80%	Information Systems
8	116	3%	2808	84%	Information systems security
9	108	3%	2916	87%	Library & information science research
10	108	3%	3024	90%	Online information review
11	87	3%	3111	93%	Journal of internet cataloging
12	70	2%	3181	95%	Information retrieval & library automation
13	67	2%	3248	97%	Knowledge organization
14	44	1%	3292	98%	Journal of Information Science and Engineering
15	34	1%	3326	99%	International forum on information and documentation
16	29	1%	3355	100%	Information retrieval
	3355	100%			

Table 5. Collection of 16 journals from the IR and related fields.

[i] http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/
[ii] http://www.rbms.nd.edu/rbms_manual/thesaurus_construction.shtml