Enriching Thesauri with Hierarchical Relationships by Pattern Matching in Dictionaries *

Lourdes Araujo and José R. Pérez-Agüera
lurdes@sip.ucm.es, jose.aguera@fdi.ucm.es

Departamento de Sistemas Informáticos y Programación.
Universidad Complutense de Madrid. Madrid 28040. Spain.

Abstract. This paper proposes a pattern matching method applied to dictionaries to identify hierarchical relationships between terms. In this work we focus on this type of relationship because we use it in the automatic generation of thesauri, which are used to improve information retrieval tasks. However the method can also be applied to identify other semantic relationships. We distinguish two kinds of patterns: structural patterns, composed of a sequence of part-of-speech tags, and key patterns, typical of dictionary entries, composed of some key terms, along with some part-of-speech tags. This kind of patterns are automatically extracted for the dictionary entries by means of stochastic techniques. The thesaurus, that has been partially constructed previously, is then extended with the new relationships obtained by applying the patterns to a dictionary. We have based the system evaluation on the results obtained with and without the thesaurus in an information retrieval task proposed by the Cross-Language Evaluation Forum (CLEF). The results of these experiments have revealed a clear improvement on the performance.

keywords: automatic thesaurus extraction, information retrieval, query expansion, pattern matching, dictionary

1 Introduction

Information retrieval (IR) techniques aim at providing fast and effective access to a large amount of information. During the last decades IR has extended its application area from textual documents in static collections to Internet and the Web. Nowadays, IR methods include document indexing, document classification and categorization, etc., most of which try to improve the response to a search query in internet, probably the task most commonly performed everywhere and everytime.

The performance of an IR system is usually proportional to the size of the query [14]. Long queries typically provide enough information for the system

* Supported by Ingeniería del Software e Inteligencia Artificial group, ref. 910494 and project TIC2003-09481-C04

to respond with appropriate documents, while short queries usually yield a low performance. In these cases query expansion can improve the retrieval performance. A common technique to expand the query adding related terms is to use thesauri. A thesaurus is a structured list of terms, usually related to a particular domain of knowledge. Thesauri are used to standardize terminology and provide alternative and preferred terms for any application. In particular, they are very useful in keyword searching on the web if they are applied to expand the list of keywords in such a way that the searched concept is given the form it really has in the web pages relevant to a searcher's area of interest.

In spite of the great interest thesaurus have reached nowadays for web applications, most of them are manually generated, what is very expensive and limits its availability to some particular topics. Furthermore, a thesaurus usually requires to be periodically updated to include new terminology, in particular in modern terms, such as those related to computer science. These reasons make the automatic generation of thesauri an interesting area of research which is attracting a lot of interest. Research on automatic thesaurus generation for information retrieval began with Sparck Jones's works on automatic term classification [10], G. Salton's work on automatic thesaurus construction and query expansion [16], and Van Rijsbergen's work on term co-occurrence [17]. Voorhees [18] applied a different approach, based on linguistic information obtained from WordNet, to perform query expansion, with very limited results. In the nineties Qui and Frei [14]. worked on a term-vs-term similarity matrix based on how the terms of the collection are indexed. Recently, Zazo, Berrocal, Figuerola and Rodríguez [2] have developed a work using similarity thesauri for Spanish documents. Jing and Croft [9] have proposed an approach to automatically construct collection-dependent association thesauri using large full-text documents collections. Those approaches obtain promising results when applied to improve information retrieval processes.

The goal of this work is to enrich the structure of a thesaurus with hierarchical relationships or taxonomy extracted from a dictionary. This is done by automatically extracting from a dictionary patterns which indicate this kind of relationship. Many entries to a typical explanatory dictionary usually adopt predefined forms. For example, let us consider some typical entries from the English online dictionary *dictionary.com*:

entry	definition
chemical	Of or relating to chemistry.
physical	Of or relating to material things...
numerical	Of or relating to a number or...

It is easy to observe patterns such as *Of or relating to NOUN* and *Of or relating to ARTICLE NOUN*. These patterns can be automatically extracted from the dictionary.

On the other hand, even dictionary entries which do not contain key expressions usually present a restricted structure which allow extracting semantic information with only a naive analysis. This property has been exploited in different

manners in research in natural language processing. Alshawi [1] applies a hierarchy of phrasal patterns to analyze dictionary word sense definitions applied to a particular dictionary: the *Longman Dictionary of Contemporary English*, which uses a restricted vocabulary in its definitions. Chodorow and Byrd [4] propose some semi-automatic procedures for extracting and organizing information implicit in dictionary definitions. The system Mindnet [15], based on the use of a broad-coverage NL-parser, has been applied to dictionary definitions. Jannink [8] uses a kind of PageRank algorithm for extraction of hierarchical relationships between words in a dictionary. Markowitz et al. [12] use dictionary patterns to find the features of a lexicon entries, such as verb categories, selection restrictions, etc. Other works search for patterns to identify semantic relationships in other resources such as large text corpora [7] and free text [11].

To search the hierarchical relationships in dictionary entries which do not contain key expressions we propose to use a collection of simple part-of-speech patterns or structural patterns. In this case the potential relationships are checked by applying vector space similarity measures on a text collection concerning the intended thesaurus domain.

The thesaurus to be enriched has been previously generated applying statistical techniques for the selection of terms and the detection of term relationships. The particular domain of knowledge to which the new thesaurus is devoted, is characterized by a set of terms extracted from a document collection about the intended topic. This is done by applying indexing techniques. Then we use the information previously collected in other thesauri about these terms to construct the initial structure of the new one. Finally, the new thesaurus is enriched by searching for new relationship among its terms. The three basic relationships between the terms of a thesaurus are equivalence (they are synonyms, one is the translation of the other, its archaic form, etc.), hierarchical and associative relationships. Equivalence is directly extracted form a dictionary, while the hierarchical relationship amounts to extracting by the pattern matching. The associative relationship between terms which are not connected by a hierarchy[1] is first detected using co-occurrence measures and, its type is characterized later.

The system has been evaluated by comparing the results obtained in an information retrieval task, for which the expected results are perfectly defined, when a set of query terms are directly consulted, and when they are previously expanded with the generated thesaurus. For the evaluation, we have used one of the document collections provided by the Cross-Language Evaluation Forum (CLEF)[2], which have been specifically developed for testing and evaluating information retrieval systems.

The rest of the paper proceeds as follows: section 2 describes the different techniques used to generate the initial structure of the thesaurus; section 3 is devoted to describe the mechanism to extract the dictionary patterns which indi-

[1] for example because they are narrower terms of different broad terms, but they still present some kind of relationship.

[2] http://www.clef-campaign.org/

cate hierarchical relationships; section 4 presents and discusses the experiments and results, and section 5 presents the main conclusions of this work.

2 First Steps in the Thesaurus Generation

In this work we combine different techniques to obtain a new thesaurus for a particular domain of knowledge.

2.1 Term Selection: the Core Set

The first step in the construction of the new thesaurus is the selection of the *core set* of terms which characterize the intended domain, according to the provided text collection. The text collection is previously preprocessed in order to determine the index terms. We perform a POS tagging of the documents to identify nouns, the words which can be included in the thesaurus. We eliminate typical stop words (articles, prepositions, conjunctions, etc). But we also eliminate other terms, that we call *specific stopwords* which are not typical stopwords, but which are too frequent in the collection to be good discriminators for thesaurus construction. Examples of specific stopwords are months, name of the days, etc. Words resulting from the previous step are applied a stemming process. The last step is the selection of the most representative terms of the text to be used as index. This phase is carried out by applying the standard indexing technique TF-IDF, i.e. the construction of an index for each document which characterizes it and allows a quicker access than the whole set of words of the document. We have used the classic inversion technique in information retrieval, constructing an inverted index whose terms have associated a list of pointers to the occurrences of the term in the text collection.

2.2 Generation of the Intersection Thesaurus

The next step of the process is the generation of the *intersection thesaurus* from a set of source thesauri, if there is any. In other case the procedure would go to the next phase. The source thesauri that we have used are the following ones:

- EUROVOC, which contains concepts on the activity of the European Union.
- SPINES, a controlled and structured vocabulary for information processing in the field of science and technology for development.
- ISOC, thesaurus aimed at the treatment of information on economy.

It is not possible to find conflict between the hierarchies provided by different sources thesauri for a same term of the core set, because the norm z39.19 allows the existence of polyhierarchical relationships [13][3]. For this reason, whenever we found two broader terms for the same term in the thesauri source, we used both in the intersection thesaurus generating its respective entries.

[3] page 18

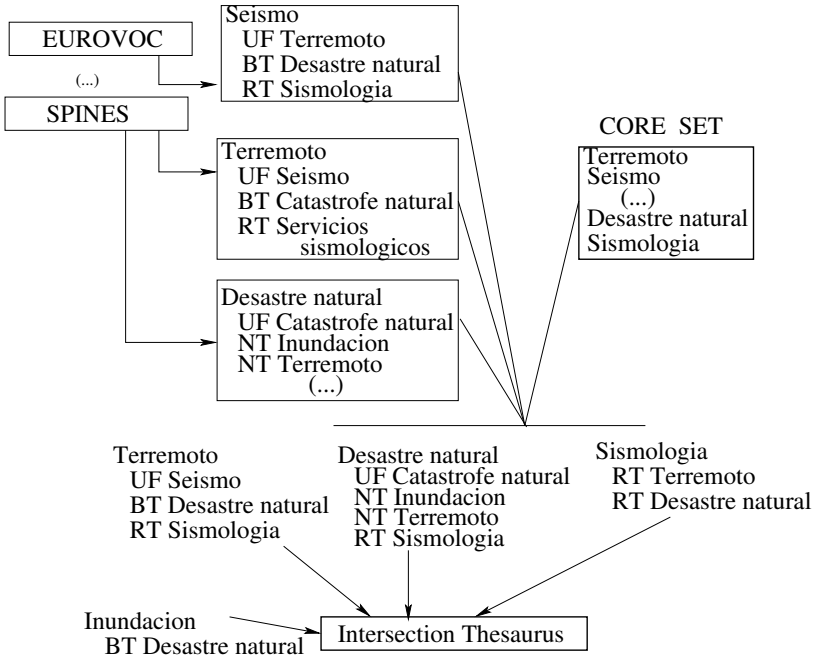

Fig. 1. Example of generation of the intersection thesaurus. UF stands for *used for*, NT for *narrower term*, BT for *broader term* and RT for *related term*.

Terms which appear in both, the core set and any source thesauri, are the term list of the intersection thesaurus. Furthermore, the relationships among the terms included in the new thesaurus are provided by the source thesauri. Figure 1 shows an example of generation of the intersection thesaurus. When the term *terremoto* (earthquake), which belongs to the core set, is searched in the source thesauri two entries are found, one in SPINES and the other one in EUROVOC.

In EUROVOC *terremoto* (earthquake) belongs to an entry whose preferred term is *seísmo* (seism) and which also contains *desastre natural* (natural disaster) (BT), and *sismología* (seismology) (RT). In SPINES *terremoto* is the preferred term of an entry which also contains the synonym *seísmo*, the broader term *catástrofe natural* (natural catastrophe) and the related term *servicios sismológicos* (seismological service). In SPINES, *terremoto* also appears in other entry whose preferred term is *desastre natural*, and which also contains the synonym *catástrofe natural*, and the narrower terms *inundación* (flood) and *terremoto*. Accordingly, the intersection thesaurus presents entries whose preferred terms are *terremoto*, *desastre natural*, *sismología* and *inundación*, the terms of the core set. *Seísmo*, which also belongs to the core set, is not given an entry because it is equivalent to *terremoto*, which appears before in the core set. Each entry is composed of the terms connected with the preferred one, or with its equivalent terms, in any of the source thesauri. Thus, the *terremoto* entry is

composed of *seísmo*, connected to *terremoto* in both thesauri, *desastre natural*, connected to *seísmo* (which is equivalent to *terremoto* in EUROVOC), etc.

2.3 Generation of the new Thesaurus

Finally, the structure of the new thesaurus is extended with new relationships among its terms. If the couple of terms to be related appears in some of the source thesauri, this indicates the kind of its relationships. If they do not appear in the source thesauri, its possible relationship has to be investigated. Each type of relationship is studied in a different manner. Equivalence is extracted from Eurowordnet [5]. Hierarchical relationships are extracted by the pattern matching techniques described in the next section. To detect associative relationships we determine the pairs of terms for which the semantic similarity is significant enough using the classic measure of cosine (the similarity is above a threshold value of 0.3 in our case).

3 Pattern Matching for Relationships Identification

Let us now consider the technique used to detect hierarchical relationships. It relies on the assumption that in a dictionary the entries for a term which is an instance of a more general concept contain a reference to the term for this general concept. Furthermore, we assume that the references to more general terms usually adopt some predefined patterns. We consider two kinds of patterns: *structural patterns*, which are defined as a sequence of part-of speech tags, and *keyphrase patterns*, which are composed of a keyphrase along with some part-of-speech tags. Each of these types is used in a different manner. Structural patterns are used to check the relationships between the expression which is the entry to the dictionary and the noun phrase expressions which appear in the definition, wherever they appear. Obviously, in general both expressions may not be related at all. Accordingly the relationship of these pairs of terms is checked in the training texts which define the domain.

We have considered the following set of structural patterns for the detection of hierarchical relationships:

noun
noun adjective
noun noun
noun preposition noun
noun preposition article noun

On the other hand, keyphrase patterns automatically identify the expression of the definition which is related with to dictionary entry and the type of this relationship. For example, let us consider the following entries from the RAE (Real Academia Española) Spanish online dictionary:

entry	definition
campesino	Perteneciente o relativo al campo.
ciudadano	Perteneciente o relativo a la ciudad.
marino	Perteneciente o relativo al mar.

It is easy to observe the pattern *perteneciente o relativo a*, which means *belonging or related to*. We have designed a method to automatically detect such keyphrases patterns in an online dictionary. Our method is as follows:

- The first step is the selection of the key terms which form the expression. A sequence of adjacent terms is considered a key expression if it is frequent enough, i.e. if the number of occurrences in the dictionary is above a threshold. Because the key expressions have very different lengths, we compute the frequency of sequences of different length (from 2 to 10). The threshold decreases with the length of the sequence, since short sequences may be frequent even if they are not a key expression.
- The previous step produces lists of key expressions of different length. In general, some key expressions will appear as part of other expressions from other lists. In these cases we must select one of them, as complete and as general as possible. If several expressions from the $ngram_i$ are part of one expression from $ngram_{i+1}$, then we select this expression because it is more complete. For example,

ngram-4: cada una de las
una de las partes
ngram-5: cada una de las partes

Our method selects the expression *cada una de las partes* from the *ngram-5* which is the most complete one.
On the other hand, if an expression from the list $ngram_i$ is part of several expressions from the list $ngram_{i+1}$, we consider that the expression of list $ngram_i$ is more general, and thus it is the one selected. For example, we have extracted the following data from the RAE (Real Academia Española) Spanish dictionary.:

ngram-4: perteneciente o relativo a
ngram-5: perteneciente o relativo a este
perteneciente o relativo a esta
perteneciente o relativo a el
perteneciente o relativo a la
perteneciente o relativo a las
perteneciente o relativo a los
perteneciente o relativo a un
perteneciente o relativo a una

Our method selects the expression *perteneciente o relativo a* from the *ngram-4* because it is more general, since it corresponds to several expression in the ngram-5 list.

– Once we have selected the key terms, patterns must be completed with the part-of-speech tags which give rise to frequent patterns. For example, the key phrases from the examples above give rise to the following patterns:

key expression	pattern
cada una de las partes de	cada una de las partes de ART N
perteneciente o relativo a	perteneciente o relativo a ART N
	perteneciente o relativo a DADJ N

where ART stands for article, N for noun and DADJ for demonstrative adjective. For a word sequence which matches one of these patterns the noun assigned to N is known to be a more general term than the word which is the entry for the dictionary.

Because patterns are dictionary dependent, they must be extracted for each dictionary we want to use. However, the described method is automatic and can be applied to any electronic dictionary and in any language. We have developed our experiments in Spanish, using the RAE (Real Academia Española) dictionary. We have performed a part-of-speech (POS) tagging of the dictionary entries in order to detect the selected structures. We have used SVMTool [6] for tagging, a software which implements a POS-Tagger with Support Vector Machine and achieves an accuracy of 96,7% for Spanish texts and 97,8% for English texts. Figure 2 shows an example of taxonomy generated using only structural patterns (a) and with both kinds of patterns (b) extracted from the RAE dictionary.

4 Experiments and Results

The prototype developed for our experiments has been implemented using the programming language Java. This prototype has been run on a computer Intel Pentium IV Hyper-Threading 3.40 GHz, with 2GB of RAM memory.

In order to provide a quantitative measure for the quality of the generated thesaurus, we have decided to evaluate its usefulness when it is applied to an information retrieval task. Specifically, we used the thesaurus to perform a term-to-term query expansion, i.e. for identifying terms related with the query terms in order to improve the retrieval capability.

For query expansion we use the method proposed by Qiu y Frei [14], which selects expansion terms according to their similarity with all query terms. Given a query q composed of terms $(t_1, t_2, ..., t_n)$ which are assigned weights $(w_1, w_2, ..., w_n)$, the similarity with a term $t\prime$ is defined as follow:

$$sim_{qt\prime}(q, t\prime) = \sum_{t_i \epsilon q} w_i * sim(t_i, t\prime) \quad (1)$$

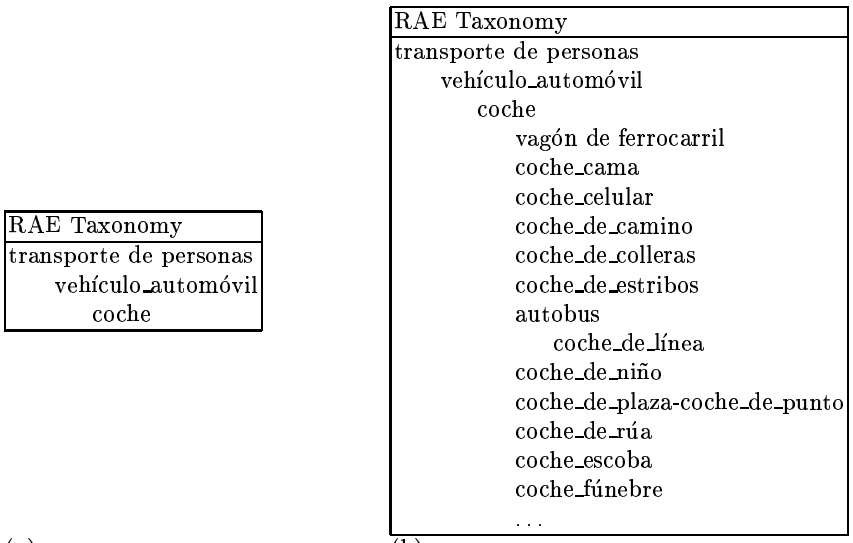

Fig. 2. Example of taxonomy generated with structural patterns (a) and with both kind of patterns (b).

where $sim(t_i, t\prime)$ is the similarity[4] computed when the thesaurus is generated. The weight of each expansion term $t\prime$ with respect to the query q is defined as:

$$w_{qexp}(q, t\prime) = \frac{sim_{qt\prime}(q, t\prime)}{\sum_{t_i \epsilon q} w_i} \tag{2}$$

This weight can be interpreted as the weighted mean of similarities between the candidate term and all terms in the query. We use this weight as a boost factor in the TF-IDF expression of our search engine[5].

With the aim at being as fair as possible, in the selection of tests we have taken a set of tests used in the CLEF (Cross-Language Evaluation Forum) for the Spanish language. The collection and tests used come from EFE94. This document collection came from the international news agency EFE, from all the news received during 1994.

For the evaluation of the system we have used the classic measures of precision and recall [3]. Recall is the fraction of the relevant documents which have been retrieved and precision is the fraction of the retrieved documents which are relevant. Specifically we use R-precision, which is the precision after retrieving R documents, where R is the total number of relevant documents for the query.

[4] It is the measured similarity for terms extracted from texts, and the similarity value of the corresponding dictionary entry for terms taken from the dictionary.

[5] We use lucene (http://lucene.apache.org/java/docs/) for the implementation of the our search engine which use Vector Space Model like retrieval paradigm

As a battery of test we have used a total of 40 extracted queries of the batteries provided by CLEF in 2001 and 2002. The reason why we have made a selection in the batteries of tests is our need of using a set of queries that can be expanded by thesauri source and whose domain is focused on politics and economy. Accordingly, we discard some queries on other topics, such as those related to sports.

Query	R-prec	R-prec improvement	Recall	Recall Improvement
Original	0.4891	–	0.61	–
Spines	0.4196	- 14.20%	0.6213	+ 1.81%
Eurovoc	0.4113	- 15.90%	0.6442	+ 5.3%
ISOC-Economy	0.4122	- 15.72%	0.6231	+ 2.1%
Intersection Thesaurus	0.3813	- 22.04%	0.6771	+ 9.9%
RAE-taxonomy-key (only key patterns)	0.4978	+ 1.74%	0.6483	+ 5.9%
RAE-taxonomy (both pattern types)	0.5116	+ 4.39%	0.6663	+ 8,44%
Final Thesaurus	0.5614	+ 12,87%	0.7312	+ 16.57%

Table 1. Precision and recall results for a set of queries from EFE94 provided by CLEF.

Table 1 shows the results obtained using different thesauri to expand a set of queries from the EFE94 collection provided by CLEF[6]. The first row corresponds to the query without expansion. The next three rows present the results expanding the query with Spines, Eurovoc and ISOC-Economy thesauri, respectively. The fifth row corresponds to an expansion with our intersection thesaurus, composed of terms from the text collection and from source thesauri. The 6th row corresponds to expanding the query with the taxonomy obtained by only applying key patterns. The 7th row presents the results expanding with the taxonomy obtained using both types of patterns. Finally, the last row gives the results expanded with our final thesaurus, which also includes equivalent and associative relationships. We can observe that recall improves in every case since the set of search terms is enlarged with thesaurus terms. Precision also improves in the last three rows because the percentage of relevant documents retrieved with the query expansion is larger than that for the original query, i.e. ambiguity has been reduced. However precision gets worse for the source thesauri and their intersection. This means that they provide too general terms for the query expansion. Results show that both, key and structural patterns, reduce ambiguity and thus improve precision.

[6] http://clef.isti.cnr.it/

5 Conclusions and Future Works

This paper describes a method to use dictionary patterns in the construction of thesauri. Dictionary patterns of different lengths are automatically extracted from the dictionary entries. We also propose a method to automatically generate the thesaurus and enrich it with the hierarchical relationships detected with the extracted patterns. This paper shows how to use handmade thesauri for the automatic generation of new thesauri. There exists a large amount of handmade thesauri and they are very useful as knowledge bases for the automatic generation of thesauri[7]. Furthermore, we have defined a methodology to combine linguistic methods and statistical methods for the automatic generation of thesauri. Results have shown the usefulness of the generated thesauri, improving both, recall and precision measures in an information retrieval task. Key patterns have been proved useful to include new terms in the thesaurus taxonomy without increasing the ambiguity because they correspond to very specific relationships. Structural patterns are also useful, but to avoid increasing ambiguity the terms selected by them are only included in the taxonomy if they belong to the text collection which defines the domain. We have used a Spanish dictionary in our experiments, but the method is valid for any language, though the dictionary patterns extracted will depend on the particular language and on the dictionary used.

For the future we expect to improve results by using more dictionaries in the process. We will also try to improve the performance of the information retrieval tasks by weighting the relationships used in the query expansion. We also plan to extend the method to detect patterns for more specific relationships, such as *be part of*, etc.

References

1. Hiyan Alshawi. Processing dictionary definitions with phrasal pattern hierarchies. *Comput. Linguist.*, 13(3-4):195–202, 1987.
2. Angel F. Zazo and Carlos G. Figuerola and Jose L. Alonso Berrocal and Emilio Rodríguez. Reformulation of queries using similarity thesauri. *Information Processing and Management*, 41(5):1163–1173, 2005.
3. Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.
4. Martin S. Chodorow, Roy J. Byrd, and George E. Heidorn. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd annual meeting on Association for Computational Linguistics*, pages 299–304, Morristown, NJ, USA, 1985. Association for Computational Linguistics.
5. P. Vossen (Ed.). *EuroWordNet A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic publishers., 1998.
6. Jesús Giménez and Lluís Márquez. Svmtool: A general pos tagger generator based on support vector machines. In *Proceedings of the 4th LREC*, 2004.

[7] Web Thesaurus Compendium: http://www.ipsi.fraunhofer.de/Ĩutes/thesoecd.html

7. Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 539–545, Morristown, NJ, USA, 1992. Association for Computational Linguistics.
8. J. Jannink and G. Wiederhold. Thesaurus entry extraction from an on-line dictionary. In *Fusion '99*, pages 110–138, 1999.
9. Y. Jing and W. Bruce Croft. An association thesaurus for information retrieval. In *Proceedings of RIAO-94, 4th International Conference "Recherche d'Information Assistee par Ordinateur"*, pages 146–160, New York, US, 1994.
10. K. Spark Jones and R.M. Needham. Automatic Term Classification and Retrival. *Information Processing and Management*, 4(1):91–100, 1968.
11. Juan Lloréns and Hernán Astudillo. Automatic generation of hierarchical taxonomies from free text using linguistic algorithms. In *OOIS Workshops*, pages 74–83, 2002.
12. Judith Markowitz, Thomas Ahlswede, and Martha Evens. Semantically significant patterns in dictionary definitions. In *Proceedings of the 24th annual meeting on Association for Computational Linguistics*, pages 112–119, Morristown, NJ, USA, 1986. Association for Computational Linguistics.
13. National Information Standards Organization (U.S.). *Guidelines for the Construction, Format, and Management of Monolingual Thesauri*, volume ANSI/NISO 239.19-1993 of *National information standards series*. NISO PRESS, 1994.
14. Yonggang Qiu and Hans-Peter Frei. Concept-based query expansion. In *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval*, pages 160–169, Pittsburgh, US, 1993.
15. Stephen D. Richardson, William B. Dolan, and Lucy Vanderwende. Mindnet: acquiring and structuring semantic information from text. In *Proceedings of the 17th international conference on Computational linguistics*, pages 1098–1102. Association for Computational Linguistics, 1998.
16. G. Salton, C. Buckley, and C. T. Yu. An evaluation of term dependence models in information retrieval. In *SIGIR '82: Proceedings of the 5th annual ACM conference on Research and development in information retrieval*, pages 151–173, New York, NY, USA, 1982. Springer-Verlag New York, Inc.
17. C.J van. Rijsbergen, D.J. Harper, and M.F. Porter. The selection of good search terms. *Information Processing and Management*, 17(2):77–91, 1981.
18. Ellen M. Voorhees. Using wordnet to disambiguate word senses for text retrieval. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 171–180, New York, NY, USA, 1993. ACM Press.