Automated Generation of Category-Specific Thesauri for Interactive Query Expansion[*]

Giuseppe Attardi and Sergio Di Marco
Dipartimento di Informatica
Università di Pisa
Corso Italia, 40
56125 Pisa (italy)
E-mail: {attardi,dimarco}@di.unipi.it

Fabrizio Sebastiani
Istituto di Elaborazione dell'Informazione
Consiglio Nazionale delle Ricerche
Via S. Maria, 46 - 56126 Pisa (Italy)
E-mail: fabrizio@iei.pi.cnr.it

Abstract

The categorisation of documents into subject-specific categories is a useful enhancement for large document collections addressed by information retrieval systems, as a user can first browse a category tree in search of the category that best matches her interests, and then issue a query for more specific documents "from within the category". This approach combines two modalities in information seeking that are most popular in Web-based search engines, i.e. category-based site browsing (as exemplified by e.g. YAHOO$^{\mathrm{TM}}$) and keyword-based document querying (as exemplified by e.g. ALTAVISTA$^{\mathrm{TM}}$). Appropriate query expansion tools need to be provided, though, in order to allow the user to incrementally refine her query through further retrieval passes, thus allowing the system to produce a series of subsequent document rankings that hopefully converge to the user's expected ranking. In this work we propose that automatically generated, category-specific "associative" thesauri be used for such purpose. We discuss a method for their generation, and discuss how the thesaurus specific to a given category may usefully be endowed with "gateways" to the thesauri specific to its parent and children categories.

Categories and subject descriptors: H.3.1 [Information storage and retrieval]: Content analysis and indexing - *Indexing methods; Thesauruses.* H.3.3 [Information storage and retrieval]: Information search and retrieval - *Query formulation.*

[*] This work has been carried out in the context of the project EUROSEARCH LE4-8303, funded by the Commission of the European Communities under the ESPRIT Telematics scheme.

1 Introduction

We here report work in progress within the EUROSEARCH project, whose purpose is the design and implementation of a European federation of n search engines $\mathcal{E}_1, \ldots, \mathcal{E}_n$, each addressing a national Web space of documents expressed in the respective languages $\mathcal{L}(\mathcal{E}_1), \ldots, \mathcal{L}(\mathcal{E}_n)$. Each search engine $\mathcal{E}_i$ in the federation will be capable of answering queries[1] q_j worded in $\mathcal{L}(\mathcal{E}_i)$ that ask for documents written in either of a set of languages $\mathcal{L}(\mathcal{E}_{1j}), \ldots, \mathcal{L}(\mathcal{E}_{mj})$ contained in $\{\mathcal{L}(\mathcal{E}_1), \ldots, \mathcal{L}(\mathcal{E}_n)\}$. The search engine will achieve this by translating, in collaboration with search engines $\mathcal{E}_{1j}, \ldots, \mathcal{E}_{mj}$, the query from $\mathcal{L}(\mathcal{E}_i)$ to each of the languages $\mathcal{L}(\mathcal{E}_{1j}), \ldots, \mathcal{L}(\mathcal{E}_{mj})$, dispatching the translated queries to the appropriate engines, and presenting to the user the results returned by them. Of course, $\mathcal{E}_i$ will also search within its own national Web space in case, as most of the times will indeed happen, $\mathcal{L}(\mathcal{E}_i) \in \{\mathcal{L}(\mathcal{E}_{1j}), \ldots, \mathcal{L}(\mathcal{E}_{mj})\}$[2].

1.1 Modalities of information seeking

One key aspect of the EUROSEARCH architecture is the combination of two information access modalities that have traditionally been addressed by separate tools.

The first modality is *keyword-based document querying*, exemplified by the ALTAVISTA™ search engine (`http://www.altavista.digital.com/`), in which a user types in a list of keywords or phrases and receives back as a result a list of documents, ranked in decreasing probability (or estimated degree) of relevance to the query, in which these keywords play a (decreasingly) significant role (e.g. appear in the title, or appear in headers, or in the body of the document). The second modality is *category-based site browsing*, exemplified by the YAHOO™ search engine (`http://www.yahoo.com/`), in which a user is allowed to browse a hierarchy of subject categories, each of which contains pointers to a *select* number of Web sites highly relevant to it; after navigating down to the category she is interested in, the user can then browse the sites referred therein.

The EUROSEARCH architecture integrates the two modalities by allowing the user to first browse a hierarchy of categories, thus docking to the category $\mathcal{C}$ she was looking for, and from there issue keyword-based queries for documents

1. *belonging* to $\mathcal{C}$; documents satisfying this *must* have been categorised under $\mathcal{C}$ by the search engine, likely because they reside on the top relevant sites for $\mathcal{C}$;

2. *related* to $\mathcal{C}$; documents satisfying this *need not* have been categorised under $\mathcal{C}$ by the search engine, but they must nevertheless be topically related to $\mathcal{C}$.

Traditional search engines based on category-based site browsing, such as YAHOO™, tend to incorporate functionality (1) but not functionality (2). We allow both (1) and (2) as alternative search modes, because while (1) yields high *precision* (i.e. a high proportion of the retrieved documents tend to be relevant to the user's need), the second yields high *recall* (i.e. a high proportion

[1] In information retrieval, a distinction is often made between the user *request*, usually a list of words or phrases in natural language by which the user expresses her information need, and the *query*, the internal representation that the system gives to the request (e.g. by weighting the individual words differently). For simplicity we will avoid this distinction here and always use the term "query", allowing context to discriminate between the two meanings from time to time.

[2] For simplicity, here we do not discuss the case in which a search engine $\mathcal{E}_i$ also addresses documents that, although residing within its own national Web space, are expressed in a language different from $\mathcal{L}(\mathcal{E}_i)$. This case is indeed tackled in EUROSEARCH, the obvious case being when this language is English.

of the documents relevant to the user's need tend to be retrieved) [1]; different users have different requirements in terms of recall and precision, and their needs are properly catered for by allowing the user to choose between the two alternatives[3].

The advantage of modality (2) over pure keyword-based document querying is that the query is less ambiguous, or more focused, than it would be if issued from a semantically neutral environment (such as AltaVista$^{\mathrm{TM}}$ or the "root" category of Yahoo$^{\mathrm{TM}}$), as terms contained in the query can appropriately be interpreted by the search engine in the semantic context of the category within which the query originated. For instance, the word `bank` contained in a query issued from within the `BusinessAndEconomy` category can be safely interpreted in its sense of "financial institution" rather than in its sense of "fortification of a river", and can thus be both interpreted correctly and translated correctly into the target languages. In Eurosearch, word sense disambiguation prior to translation into the other languages is of fundamental importance, especially as in the Eurosearch architecture a "pivot" language (English) is used[4]: word sense ambiguities rooted in the pivot language would not only degrade the performance, but cause results unintelligible to a user not proficient in the pivot language.

The profile of the category (typically, a vector of weighted terms that are deemed relevant to it) can be used as contextual information; the net effect is one of word sense disambiguation, both when the query is addressed to the local search engine (this is the case in which $\mathcal{L}(\mathcal{E}_i) \in \{\mathcal{L}(\mathcal{E}_{1j}), \ldots, \mathcal{L}(\mathcal{E}_{mj})\}$) and when it is translated into the pivot language in order to be dispatched to other search engines. The typically large size of the category profile usually ensures high-quality disambiguation [18].

1.2 Query refinement

No matter whether they are issued from within categories or from semantically neutral environments, queries always return less than perfect results: some irrelevant documents are ranked high in the list, and some relevant documents are ranked low. It is thus necessary to offer the user tools for *query refinement*, by means of which she may interact with the system and feed it information that allows it to refine the query through further retrieval passes, thus yielding a series of subsequent document rankings that hopefully converge to the user's expected ranking. Modern query refinement methods may loosely be classified into *automatic query expansion* (AQE) methods and *interactive query expansion* (IQE) ones (see e.g. [5]).

Many AQE methods rely on *relevance feedback* techniques (see e.g. [17]), whereby terms featuring prominently in documents marked relevant by the user are automatically added to the query (or reweighted, increasing their weight appropriately); other AQE methods rely instead on the use of *thesauri*, from which terms semantically related to the query terms [7] or to the query as a whole [13] are extracted and added to the query. These techniques have often given good

[3] A manager who wants a quick introduction to a novel topic is a typical precision-oriented user; she wants just a few highly relevant articles, and she wants them rightaway. On the contrary, an author who is preparing a review article on a given subject is a typical recall-oriented user: she wants to be certain that she reviews most or possibly all of the relevant material, and in order to do this she is also prepared to shuffle through some possibly irrelevant material. See [10] for a more thorough discussion on this.

[4] This means that a query issued in Italian and asking for documents in Spanish is translated from Italian to English by the Italian engine, and then dispatched to the Spanish engine, which receives it and translates it from English to Spanish. This not only allows to cut down on the number of required bilingual dictionaries from $O(n^2)$ to $O(n)$, but also solves the problem engendered by the non-availability of an online bilingual dictionary for a given pair of languages. See [?] for details.

results, but their effectiveness is known to depend on a lot of factors, including the method used for document ranking, the characteristics of the document collection being targeted, and the number of terms used.

IQE methods (also called *semi-automatic query expansion methods* [13]) are instead based on the idea that these automatically selected terms must be first submitted to the user (by a *term suggestion device* [19]), who may thus decide which to include and which not to include in the revised query[5]. Again, the suggested terms may come from relevance feedback, or from lexical resources such as thesauri. While in the first case the results reported in the literature are mixed, showing that only experienced users tend to benefit from these techniques [11], in the latter case there have been promising results, that have prompted a flurry of work in the area (see e.g. [4, 19]), especially within the context of the Digital Libraries Initiative.

Querying from within categories also requires that query refinement too be performed in a category-specific way. In the rest of the paper we discuss work in progress within EUROSEARCH aimed at allowing the user to interactively add new keywords (or substitute previously used keywords with more specific ones) by choosing them from a category-specific *associative thesaurus*, i.e. a graph in which nodes represent terms and edges represent (unspecific) relationships of semantic similarity between terms. The advantage of associative thesauri over other kinds of thesauri (such as conventional "hierarchical" ones) is that they may be generated automatically through statistical analysis of word occurrences in a given collection.

The rest of the paper is organised as follows. In Section 2 we discuss associative thesauri and the problem of their automatic generation. Section 3 discusses the methods we intend to follow for allowing the users to exploit the generated thesauri for interactive query refinement.

2 Associative thesauri and their automatic generation

Thesauri have long been used for query expansion or reformulation (see e.g. [5]). A *thesaurus* is a collection of words organised according to a topology that reflects the semantics of the terms and of their associations; typically, thesauri are *domain-specific*, i.e. the set of words they contain are those specific to a given discipline. In a conventional thesaurus [6] words are organized into binary relations, the most important of which are

- $NT(t_1, t_2)$ ("Narrower Term"), meaning that t_1 is a more specific term than t_2;
- $RT(t_1, t_2)$ ("Related Term"), meaning that t_1 is semantically related, albeit in a non-hierarchical way, to t_2.

The NT relation thus induces a partial order on the set of terms, giving it the characteristic hierarchical structure. This structure makes it easier for users to *substitute* terms previously used in the query by means of more specific ones, which the user can identify by browsing the thesaurus, starting from the term to be substituted and working downwards along the hierarchy.

However, conventional thesauri have been found unsuitable for suggesting to the user new terms to *add* to a query [19]. The main reason is that these thesauri are built *manually*, by expert lexicographers, and the manual nature of this process makes it virtually impossible to identify,

[5] AQE and IQE do not exclude each other: a possible strategy may be to allow both possibilities, so that the user may e.g. stay with AQE for the first retrieval passes and step in with IQE once she sees that AQE is not improving the ranking any more.

and list by means of the *RT* relation, a sufficiently large set of semantically related terms among which the user might want to choose new terms to add to the query.

A way out of this inadequacy is the *automatic* identification of semantic relationships between terms from a corpus of documents; this process originates an *associative thesaurus*[6], i.e. a graph in which nodes represent terms and edges represent relationships of semantic similarity between terms. Edges have an associated weight, denoting the strength of association between the two words; therefore, to each term is associated a list of related terms, ranked in decreasing order of association strength.

The automatic generation of associative thesauri, or variants thereof, has a long history, dating back at the very least to the seminal studies of Sparck Jones [22] and Salton [14]. More or less, the process of generating a thesaurus is articulated in the following steps:

1. standard (automatic) document indexing is performed, thereby generating the usual term-document incidence matrix that specifies a weight w_{ij} for each pair $\langle t_i, d_j \rangle$ constituted by a term t_i and a document d_j;

2. a term-term matrix is generated, specifying a "semantic relatedness" value r_{ij} for each pair of terms $\langle t_i, t_j \rangle$. The matrix may be symmetric or not, depending on the underlying notion of similarity, and is generated by taking into account factors such as the degree of co-occurrence or co-absence of the two terms in the collection;

3. small values of r_{ij} are replaced by 0, thus leaving only highly related pairs with a nonzero coefficient; these latter, which determine the edges of the resulting associative thesaurus, may be identified:

   (a) as the top n coefficients for any given term, for a pre-specified value of n;

   (b) as all the coefficients exceeding a pre-specified threshold value.

From step (2), it is clear that the resulting thesaurus embodies an *extensional*, rather than the standard *intensional*, notion of lexical semantics (see e.g. [8, page 246]): *the meaning of words is only determined by the documents they appear in.* It is also a collection-dependent notion of meaning, as the r_{ij} values strongly depend on the characteristics of the collection which is chosen as "training sample".

2.1 The EUROSEARCH approach

In a space of documents categorised into one or more categories, such as the one EUROSEARCH deals with, good terms for query expansion are most likely to be specific to the particular domain the category is about. Therefore, our approach contemplates generating *category-specific associative thesauri*, one for each category in the categorisation scheme. This generation will take place from a training set of documents previously categorised under the category of interest.

One key observation for this task is that we want to avoid pairs of terms that, although they might be deemed as strongly related in the training sample, are nevertheless extraneous to the domain-specific terminology of the category of interest. For instance, from a training sample of documents previously categorised under the `BusinessAndEconomy` label, the two words `banana`

[6] Associative thesauri are also called *term-term relationship matrices* [15], *concept spaces* [3, 4], or *similarity thesauri* [13, 20, 21].

and `coconut` might (correctly!) emerge as strongly related, simply because they co-occur in a few documents (e.g. related to stock prices of exotic fruits) and they are co-absent in most of the others. In order to avoid this, before proceeding to term-term similarity computation, we first want to identify the terms that are specific to the category of interest. These can be identified as the terms whose *within-category inverse document frequency* $WCaIDF$ is substantially smaller (i.e. smaller at least by a pre-determined factor) than their *within-collection inverse document frequency* $WCoIDF$. If Ca and Co denote the category of interest and the collection (i.e. the whole corpus), respectively, $WCaIDF$ and $WCoIDF$ may be defined as[7]

$$\begin{aligned} WCaIDF(t_i) &= log(\frac{1+\|Ca\|}{1+\#_{Ca}(t_i,d_j)}) && (1)\\ WCoIDF(t_i) &= log(\frac{1+\|Co\|}{1+\#_{Co}(t_i,d_j)}) && (2)\end{aligned}$$

where $\#_S(t_i,d_j)$ denotes the number of documents $d_j \in S$ in which term t_i occurs at least once and $\|S\|$ denotes the cardinality of set S. This criterion is based on the intuition that terms specific to a given category occur more frequently in documents belonging to the category than in "generic" documents belonging to the entire training set.

Once terms specific to the category of interest have been identified, semantic relatedness values between each pair of such terms may be computed. As mentioned in point (2) of Section 2, co-occurrence and co-absence considerations are often taken into account in this computation. We instead intend to rely on a technique developed in [20], and subsequently experimented with in [13, 21], that relies on an inversion of the roles that documents and terms traditionally have in information retrieval.

This approach may be better understood by recurring to a sort of *abstract indexing theory*, according to which a set of *items* forming a collection I are indexed by a set of *indexing features* F. Let T be the set of all *tokens* t, i.e. of all occurrences of an indexing feature f_r in an item i_s for some r, s. Let $\mathcal{F} : T \rightarrow F$ be the function that maps a token into the indexing feature of which it is an occurrence, and let $\mathcal{I} : T \rightarrow I$ be the function that maps a token into the item in which it occurs. We may thus define the *within-item frequency* of feature f_r in item i_s as the number of times f_r occurs in i_s:

$$WIF(f_r,i_s) = \|\{t \in T : (\mathcal{F}(t)=f_r) \wedge (\mathcal{I}(t)=i_s)\}\| \qquad (3)$$

We may also define the *within-collection frequency* of feature f_r as the number of items in which f_r occurs at least once:

$$WCF(f_r) = \|\{i \in I : \exists t.(\mathcal{F}(t)=f_r) \wedge (\mathcal{I}(t)=i)\}\| \qquad (4)$$

The *inverse within-collection frequency* of feature f_r is defined as:

$$IWCF(f_r) = log(\frac{1+\|I\|}{1+WCF(f_r)}) \qquad (5)$$

A possible weight for indexing feature f_r in item i_s may thus be

$$w_{rs} = WIF(f_r,i_s) \cdot IWCF(f_r) \qquad (6)$$

[7] See e.g. [16] for a clear discussion of IDF and other term weighting strategies.

If the cosine measure (i.e., inner product with cosine normalisation) is used for modelling item-item similarity, then we have

$$SIM(i_s, i_t) = \frac{\sum_{r=1}^{\|F\|} w_{rs} \cdot w_{rt}}{\sqrt{\sum_{r=1}^{\|F\|} w_{rs}^2 \cdot w_{rt}^2}} \tag{7}$$

In standard IR, documents play the role of items, terms play the role of indexing features, (3), (4) and (5) are the well-known *term frequency*, *document frequency* and *inverse document frequency* functions, respectively, while (6) is the widely used $tf * idf$ weighting scheme [16].

For the generation of the associative thesaurus, instead, *terms play the role of items and documents play the role of indexing features*; in other words, the method is simply based on the idea of "thinking dually". Term t_j is thus represented as a vector $t_j = \langle w_{1j}, \ldots, w_{nj} \rangle$, where n is the cardinality of the document collection D and w_{ij} is the weight of document d_i for term t_j. This gives rise to a "dual" vector space model, in which terms are vectors in a space generated by the documents in the collection. In this way, the most similar terms to a given term t_j may be identified by issuing t_j as a "query" term and considering the top-ranked "retrieved" terms.

Adopting this simple idea has a number of consequences, first and foremost that the intuitions that underlie the standard application to IR of this "abstract indexing theory" may be applied, *mutatis mutandis*, to the new setting. In particular, it is interesting to note that, while the rationale of $IWCF$ in the standard interpretation (i.e. *idf*) is that terms that occur in fewer documents are more valuable indexing features than the ones occurring in many documents, the rationale of $IWCF$ in the dual interpretation is that documents consisting of fewer terms (i.e. shorter documents) are likewise more valuable indexing features: shorter documents tend to deal with less topics, and hence represent the terms that occur in them in a more significant way [21].

It is clear from this description that the process of associative thesaurus generation is computationally expensive. While the term-document incidence matrix may be considered as given (as it will already have been produced in indexing the documents for "standard" retrieval purposes), formula (3) needs to be calculated for every $\langle f_r, i_s \rangle$ document-term pair, formula (4) needs to be calculated for every document f_r, etc. However, the fact that only terms specific to the category, and only documents categorised under the category, are used, substantially cuts down the complexity of the process with respect to the standard, highly expensive problem of generating an *unspecific* (i.e. topic-neutral) associative thesaurus [4]. This is true even once we consider that the process must be repeated for every single category, as this brings about only an additive increase in complexity, while the above-mentioned reduction in complexity is multiplicative. Besides, one should not forget that, as recalled in [13], the generation of the thesaurus is done once for all.

3 Using category-specific associative thesauri for query refinement

3.1 Thesaurus display and navigation

We are currently evaluating different strategies for allowing the user to refine or expand her query by browsing the thesaurus through a graphical interface. One possibility is the adoption of a graph browser [9]. Such a tool would display, upon clicking on a term in the previously issued query window, a star-shaped graph representing the portion of the thesaurus consisting of the clicked word and its semantically most related words, linked to it by edges; navigation in the graph would allow the user to select new words for addition to the query or for substitution (refinement) of the clicked word (standard techniques of graphical interfaces may adopted both for distinguishing between these two types of events and for implementing other tasks mentioned below). Various visualisation techniques may be employed here, among which adaptations of those developed within the "ostensive model of information needs" [2].

A possible alternative that we are also considering is the use of an even simpler thesaurus display technique based on hierarchical menus. In this case, clicking on a word appearing in the previous query window would result in the popping up of a menu consisting in the ranked list of the terms semantically related to the word, listed in decreasing order of strength of relatedness. This menu would be hierarchical, thus allowing the selection of a word several steps away in a single click (and, thanks to the ranking of the list, with likely minimum mouse travelling distance). One advantage of this technique over the previously discussed one, apart from the possible reduction of screen clutter, is that graph displays are clumsy for the representation of *non-symmetric* binary relationships, and the binary notion of semantic relatedness (or similarity) seems inherently to be such [12].

3.2 Jumping across inter-thesauri borders

While the methods described in Section 2.1 should ensure that most terms specific to a given category $\mathcal{C}$ have been captured within the thesaurus $\mathcal{T}(\mathcal{C})$ specific to $\mathcal{C}$, it is quite possible that some of them have been missed. However, some of the missed terms may well have been captured within the thesaurus $\mathcal{T}(\mathcal{C}')$ specific to the category $\mathcal{C}'$ that is the parent of $\mathcal{C}$ in the category tree, or within the thesaurus $\mathcal{T}(\mathcal{C}_i)$ specific to one of the categories $\mathcal{C}_1, \ldots, \mathcal{C}_k$ that are the children of $\mathcal{C}$ in the same tree. For instance, suppose that

`BusinessAndEconomy/FinanceAndInvestments/MutualFunds/IndividualFunds`

represents a branch in the category tree. The term "`fund screening tools`" might have been captured within the thesaurus of category `MutualFunds` and not in that of categories `IndividualFunds` and `FinanceAndInvestments`; nevertheless, it might well be of interest to a user issuing a query from within either of these latter categories. Although this is in principle always possible (additionally to thesaurus browsing, a user may add new terms by simply typing them in), it is indeed useful to allow the user to cross inter-thesauri borders when the involved thesauri refer to categories that stand in a parent-child relation in the category tree.

While other methods might in principle be used, we plan to use terms common to both thesauri as "gateways". In fact, the methods outlined in Section 2.1 do allow (and this will indeed be the case for several important terms) that a given term t be included in the two thesauri $\mathcal{T}(\mathcal{C})$ and $\mathcal{T}(\mathcal{C}')$, where $\mathcal{C}$ is the parent (or child) of $\mathcal{C}'$. Now, suppose that, while browsing $\mathcal{T}(\mathcal{C})$, term t is reached; along with the ranked list of terms related to t in $\mathcal{T}(\mathcal{C})$, also the list of terms related to t in $\mathcal{T}(\mathcal{C}')$ may be displayed (differentiating it, quite obviously, from the previous list by means of some adequate visualisation device). This allow the user, if she wishes so, to enter the $\mathcal{T}(\mathcal{C}')$ thesaurus and select terms from therein, notwithstanding the fact that she had started her browsing activity within a different thesaurus $\mathcal{T}(\mathcal{C})$. Jumping between more than one level of the tree is obviously made possible by iterating this strategy.

3.3 Interacting term suggestion devices

As recalled in Section 2, Schatz and his colleagues [19] have recently pointed out that, while associative thesauri are indeed useful for suggesting terms to be added to the query (*query expansion*), they are not for suggesting terms to be substituted to previously used, more generic ones (*query refinement*). The reason of this is that the relation of semantic relatedness between terms that associative thesauri encode is *untagged*, i.e. is the result of collapsing into a single relation the NT an the RT relations of conventional thesauri. The notion of term specificity needed for query refinement is thus embodied not in an associative thesaurus, but in a conventional one. Schatz and colleagues have thus proposed that two term suggestion devices (a conventional thesaurus and an associative one) be made available to the user, who may then freely jump from one to the other, thus freely intermixing query refinement and expansion.

While this strategy is indeed attractive, it has the obvious drawback that a conventional thesaurus specific to a given discipline is neither always available, nor can be generated automatically with the ease of an associative one.

However, we are considering exploiting this strategy by *allowing the category tree itself to play the role of the conventional thesaurus.* In fact, there are numerous points in favour of this decision. First, the category tree has, as the name implies, a hierarchical structure, similar to that induced by the NT relation on a conventional thesaurus. Second, the RT relation of the conventional thesaurus, that has no analogue in the category tree, is not exploited in this strategy. Third, the YAHOO^{TM} category tree we will rely on exhibits sufficient structure and depth to be used for this purpose[8].

References

[1] M. Buckland and F. Gey. The relationship between recall and precision. *Journal of the American Society for Information Science*, 45(1):12–19, 1994.

[2] I. Campbell and C. J. van Rijsbergen. The ostensive model of developing information needs. In *Proceedings of COLIS-96, 2nd International Conference on Conceptions of Library Science*, pages 251–268, Kobenhavn, DK, 1996.

[3] H. Chen and T. D. Ng. An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): symbolic branch-and-bound search vs. connectionist Hopfield net activation. *Journal of the American Society for Information Science*, 46(5):348–369, 1995.

[4] H. Chen, B. R. Schatz, T. D. Ng, J. Martinez, A. Kirchoff, and C. Lin. A parallel computing approach to creating engineering concept spaces for semantic retrieval: the Illinois Digital Library Initiative project. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):771–782, 1996.

[5] E. N. Efthimiadis. Query expansion. In M. E. Williams, editor, *Annual Review of Information Science and Technology*, volume 31, pages 121–187. American Society for Information Science, 1996.

[8] For instance, its `ComputersAndInternet` category alone contains 2533 unique subcategories distributed over 7 levels.

[6] D. J. Foskett. Thesaurus. In A. Kent, H. Lancour, and J. Daily, editors, *Encyclopedia of library and information science*, volume 30, pages 416–462. Marcel Dekker, New York, US, 1980. Also reprinted in [23], pp. 111–134.

[7] G. Grefenstette. Use of syntactic context to produce term association lists for retrieval. In *Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval*, pages 89–98, Kobenhavn, DK, 1992.

[8] S. Haack. *Philosophy of logics*. Cambridge University Press, Cambridge, UK, 1978.

[9] S. Jones, M. Gatford, S. Robertson, M. Hancock-Beaulieu, J. Secker, and S. Walker. Interactive thesaurus navigation: Intelligence rules OK? *Journal of the American Society for Information Science*, 46(1):52–59, 1995.

[10] F. W. Lancaster. Evaluation within the environment of an operating information service. In K. S. Jones, editor, *Information retrieval experiment*, pages 105–127. Butterworths, London, UK, 1981.

[11] M. Magennis and C. J. van Rijsbergen. The potential and actual effectiveness of interactive query expansion. In *Proceedings of SIGIR-97, 20th ACM International Conference on Research and Development in Information Retrieval*, pages 324–332, Philadelphia, US, 1997.

[12] H. J. Peat and P. Willett. The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42(5):378–383, 1991.

[13] Y. Qiu and H.-P. 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.

[14] G. Salton. Experiments in automatic thesaurus construction for information retrieval. In *Proceedings of the IFIP Congress*, volume TA-2, pages 43–49, Ljubljana, YU, 1971.

[15] G. Salton. Automatic term class construction using relevance. A summary of work in automatic pseudoclassification. *Information processing and management*, 16:1–15, 1980.

[16] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing and management*, 24:513–523, 1988. Also reprinted in [23], pp. 323–328.

[17] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41:288–297, 1990. Also reprinted in [23], pp. 355–364.

[18] M. Sanderson. Word sense disambiguation and information retrieval. In *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*, pages 49–57, Dublin, IE, 1994.

[19] B. R. Schatz, E. H. Johnson, P. A. Cochrane, and H. Chen. Interactive term suggestion for users of digital libraries: Using subject thesauri and co-occurrence lists for information retrieval. In *Proceedings of the 1st ACM Digital Library Conference*, pages 126–133, Bethesda, US, 1996.

[20] P. Schäuble and D. Knaus. The various roles of information structures. In *Proceedings of the 16. Jahrestagung der Gesellschaft für Klassifikation*, pages 282–290, Dortmund, DE, 1992.

[21] P. Sheridan, M. Braschler, and P. Schäuble. Cross-language information retrieval in a multilingual legal domain. In C. Peters and C. Thanos, editors, *Proceedings of ECDL-97, 1st European Conference on Research and Advanced Technology for Digital Libraries*, pages 253–268, Pisa, IT, 1997. Lecture Notes in Computer Science, number 1324, Springer Verlag, Heidelberg, DE.

[22] K. Sparck Jones. *Automatic keyword classification for information retrieval.* Butterworths, London, UK, 1971.

[23] K. Sparck Jones and P. Willett, editors. *Readings in information retrieval.* Morgan Kaufmann, San Mateo, US, 1997.