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

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

1 Introduction

The categorisation of documents into subject-specific categories is a useful enhancement for large document collections addressed by information retrieval (IR) 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!^{TM}) and keyword-based document querying (as exemplified by e.g. ALTAVISTA^{TM}). In the framework of the EUROSEARCH Project, we are addressing the problem of the *automatic* categorisation of Web documents and sites within YAHOO!-like hierarchies of categories [1,4]. The tools resulting from this project allow to overcome a major bottleneck in today's Web information organisation, i.e. the need for manual categorisation of Web documents and sites; this latter modality is inadequate, in view of the ever increasing size of the Web and of its ever evolving content.

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 *interactive query expansion* (IQE) methods are based on the idea that new terms semantically related to those present in the query should be automatically selected and submitted to the user, who may then decide which to include and which not to include in the revised query. There have been promising results in the use of IQE techniques in which the suggested terms come from lexical resources such as thesauri; these results have prompted a flurry of work in the area (see e.g. [3,6]), especially within the context of the Digital Libraries Initiative.

It is our contention that querying from within categories *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 generic relationships of semantic similarity between terms. The advantage of associative thesauri over conventional "hierarchical" ones is that they may be generated automatically through statistical analysis of word occurrences in a given collection. We discuss a method for the generation of

* 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.

category-specific associative thesauri, 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. See the full paper [8] for more details.

2 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 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` 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 is substantially smaller (i.e. smaller at least by a pre-determined factor) than their within-collection inverse document frequency. 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. For this, we rely on a technique developed in [7] that relies on an inversion of the roles that documents and terms traditionally have in IR.

2.1 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. 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 [5].

While the methods described above 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 use terms common to both thesauri as "gateways". In fact, the methods outlined above 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.

References

1. G. Attardi, A. Gullì, and F. Sebastiani. Theseus: categorization by context. In C. Hutchison and G. Lanzarone, editors, *Proceedings of THAI'99, European Symposium on Telematics, Hypermedia and Artificial Intelligence*, Varese, IT, 1999. Forthcoming.
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, 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.
4. N. Fuhr, N. Gövert, M. Lalmas, and F. Sebastiani. Categorisation tool: Final prototype. Deliverable 4.3, TELEMATICS Project LE4-8303 "EUROSEARCH", Commission of the European Communities, 1999.
5. 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.
6. 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.
7. P. Schäuble and D. Knaus. The various roles of information structures. In O. Opitz, B. Lausen, and R. Klar, editors, *Proceedings of the 16th Annual Conference of the Gesellschaft für Klassifikation*, pages 282–290. Dortmund, DE, 1992. Published by Springer Verlag, Heidelberg, DE, 1993.
8. F. Sebastiani. Automated generation of category-specific thesauri for interactive query expansion. Technical Report IEI-B4-06-1998, Istituto di Elaborazione dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, IT, 1998.