Automatic Selection of Noun Phrases as Document Descriptors in an FCA-Based Information Retrieval System

Juan M. Cigarrán, Anselmo Peñas, Julio Gonzalo, and Felisa Verdejo

Dept. Lenguajes y Sistemas Informáticos**
E.T.S.I. Informática, UNED, Madrid Spain
{juanci, anselmo, julio, felisa}@lsi.uned.es
http://nlp.uned.es

Abstract. Automatic attribute selection is a critical step when using Formal Concept Analysis (FCA) in a free text document retrieval framework. Optimal attributes as document descriptors should produce smaller, clearer and more browsable concept lattices with better clustering features. In this paper we focus on the automatic selection of noun phrases as document descriptors to build an FCA-based IR framework. We present three different phrase selection strategies which are evaluated using the *Lattice Distillation Factor* and the *Minimal Browsing Area* evaluation measures. Noun phrases are shown to produce lattices with good clustering properties, with the advantage (over simple terms) of being better intensional descriptors from the user's point of view.

1 Introduction

The main goal of an Information Retrieval (IR) system is to ease information access tasks over large document collections. Starting from a user's query, usually made in natural language, a classic IR system retrieves the set of documents relevant to the user needs and shows them using ranked lists (e.g. Google, Yahoo or Altavista).

The use of ranked lists, however, does not always satisfy the user's information needs. Ranked lists are best suitable when users know exactly what they are looking for and how to express it using the right words (e.g. the last driver for a specific graphics card or the papers published by any author). More generally, ranked lists can be useful when the task is to retrieve a very small number of relevant items. However, when there is a need to retrieve relevant information from many sources, or when the query involves fuzzy or polysemous terms, the use of a ranked list implies to read almost the whole list to find the maximum

** This work has been partially supported by the Spanish Ministry of Science and Technology within the following projects: TIC-2003-07158-C04 Answer Retrieval from Digital Documents, R2D2; and TIC-2003-07158-C04-02 Multilingual Answer Retrieval Systems and Evaluation, SyEMBRA.

B. Ganter and R. Godin (Eds.): ICFCA 2005, LNAI 3403, pp. 49–63, 2005.

number of relevant documents. For instance, if we ask *Google* (*www.google.com*) with the query '*jaguar*' looking for documents related with the jaguar as animal, we obtain 7.420.000 of web pages as a result. Of course, not all the retrieved pages are relevant to our needs and, based on the ranking algorithm of Google [1], pages containing the term '*jaguar*' but with different senses (i.e. jaguar as a car brand or jaguar as a Mac operating system) are mixed up in the resulting ranking, making the information access task tedious and time consuming.

As an alternative, clustering techniques organize search results allowing a quick focus on specific document groups and improving, as a consequence, the final precision of the system from a user's perspective. In this way, some commercial search engines (i.e. www.vivisimo.com) apply clustering to a small set of documents obtained as a result of a query or a filtering profile. The use of clustering as a post-search process applied only to a subset of the whole document collection makes clustering an enabling search technology halfway between browsing (i.e. as in web directories) and pure querying (i.e. as in Google or Yahoo).

We propose the use of Formal Concept Analysis (FCA) as an alternative to classic document clustering, not only considered as an information organization mechanism but also as a tool to drive the user's query refinement process. Advantages of FCA over standard document clustering algorithms are: a) FCA provides an intensional description of each document cluster that can be used for query modification or refinement, making groups more interpretable; and b) the clustering organization is a lattice, rather then a hierarchy, which is more natural when multiple classification is possible, and facilitates recovering from bad decisions while browsing the lattice to find relevant information.

The main drawbacks of FCA disappear when dealing with small contexts (i.e. with a small set of documents obtained as the result of a search process): a) FCA is computationally more costly than standard clustering, but when it is applied to small sets of documents (i.e. in the range of 50 to 500 documents) is efficient enough for online applications; and b) lattices generated by FCA usually are big, complex and hence difficult to use for practical browsing purposes. Again, this should not be a critical problem when the set of documents and descriptors are restricted in size by a previous search over the full document collection.

But the use of FCA for clustering the results of a free text search is not a straightforward application of FCA. Most Information Retrieval applications of FCA are domain-specific, and rely on thesauruses or (usually hierarchical) sets of keywords which cover the domain and are manually assigned as document descriptors [12, 6, 7, 5, 8, 16, 9]. The viability of using FCA in this scenario implies to solve some challenges related with: a) the automatic selection of suitable descriptors for context building, b) the rendering of node descriptions, c) the visualization of concept lattices obtained, and; d) the definition of suitable query refinement tasks. Most importantly, the (non-trivial) issue of how to evaluate and compare different approaches has barely been discussed in the past.

This paper is presented as a continuation of the research presented in [4], where the problem of automatic selection of descriptors was first addressed. In