Analysis and Validation of Information Access Through Mono, Multidimensional and Dynamic Taxonomies

Giovanni Maria Sacco

Dipartimento di Informatica, Università di Torino,
Corso Svizzera 185, 10149 Torino, Italy
sacco@di.unito.it

Abstract. Access to complex information bases through multidimensional, dynamic taxonomies (also improperly known as faceted classification systems) is rapidly becoming pervasive in industry, especially in e-commerce. In this paper, the major shortcomings of conventional, monodimensional taxonomic approaches, such as the independence of different branches of the taxonomy and insufficient scalability, are discussed. The dynamic taxonomy approach, the first and most complete model for multidimensional taxonomic access to date, is reviewed and compared to conventional taxonomies. We analyze the reducing power of dynamic taxonomies and conventional taxonomies and report experimental results on real data, which confirm that monodimensional taxonomies are not useful for browsing/retrieval on large databases, whereas dynamic taxonomies can effectively manage very large databases and exhibit a very fast convergence.

1 Introduction

Taxonomies are a well-known approach to model complex information. A taxonomy is a hierarchy of concepts going from the most general to the most specific and naturally models IS-A relationships among concepts. An example of taxonomies is Linnaean life (botanical, animal) taxonomies. In classical taxonomies, such as Linnaeus', an item (e.g. a dog) is classified under one and only one concept. The taxonomy can be seen as an efficient encoding of properties, which concisely states the properties of an item that can be recovered by following the path from the father concept to the root concept and accumulating properties: thus, we can state that a dog is a mammalian, has a spinal cord, is an animal.

Taxonomies also organize data in such a way as to orient the search for specific items: in this case, we start from the root concept and iteratively discriminate among son concepts, in order to find the appropriate one. Each time we select a concept for further expansion we reduce the number of items that we have to consider, since the items classified under the descendants of the concepts we discarded need not be considered. Thus, taxonomies can be seen as a search device that iteratively reduces the number of documents to be manually inspected by the user. Clearly, reduction can go on only up to terminal concepts: a terminal concept is not further specializable, and therefore all the documents classified under it must be manually inspected.

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 659 – 670, 2006.

Dynamic taxonomies [5, 6, 7] have been proposed as a model to describe and access large, complex information bases. From the modeling perspective, the main difference between conventional and dynamic taxonomies is that conventional taxonomies are monodimensional (i.e. an item is classified under one and only one concept), whereas dynamic taxonomies are multidimensional (i.e. an item can be classified under several different concepts at any level of abstraction). From the information access point of view, the difference between the two models is dramatic. In conventional taxonomies, the user can only iteratively expand a concept into its sons, up to the terminal level. In dynamic taxonomies, the user directly manipulates the taxonomy in order to set a focus on one or more concepts. The system is able to infer relationships between concepts and the current focus on the basis of the actual classification: related concepts are shown to the user through a reduced taxonomy, from which unrelated concepts are pruned. Concepts in the reduced taxonomy are used to effectively guide the user to reach his goal: they are all and only the concepts that can be used to set additional foci and allow further refinements even for terminal concepts.

This type of interaction implements a new paradigm of access that we call "**guided information thinning**". This paradigm can be seen as an iterative combination of exploring the information base in order to find relevant features, selecting one or more features of interest and discarding all the items that do not have these features in order to reduce the number of candidates. Subsequently, the candidates are summarized through a reduced taxonomy, from which other concepts can be selected to refine the current focus, and so on. Obviously, only a single iteration can occur in conventional taxonomies. We contend that the vast majority of user access to information benefits from this paradigm, because it is not based on precise queries (e.g. how many suppliers supply red parts…) but rather requires the identification of a set of candidate items according to specifications that are often imprecise or often not known beforehand. We further suggest that the current widespread feeling that search "does not work" is mainly due to the fact that different, ineffective paradigms (such as database queries or text retrieval) are used instead.

We want to stress here that interactive end-user access to complex information bases requires a holistic approach, in which data modeling issues are considered at least as important as human interaction ones and, in fact, data modeling and manipulation primitives can be supported only if they are easily understood by end-users. The present approach is therefore quite different in philosophy and application from most current research on ontologies and Semantic Web, which consider expressiveness and semantic richness as their main goals. Although more powerful and expressive than dynamic taxonomies, general semantic schemata are difficult to understand and manipulate by the casual user. They are better suited to programmatic access and user interaction must be mediated by specialized agents. This increases costs, time to market and decreases generality and flexibility of user access.

From this point of view, we believe that an important result for dynamic taxonomies is that a very effective and general visual exploration framework can be derived, based on conceptual machinery that is minimal and easily understood by users. In addition to the rapidly growing adoption of systems based on dynamic taxonomies by major players such as Yahoo!, Kelkoo, Bizrate, etc. starting in late 2004, Hearst et al. [2, 12] reported usability studies for systems based on dynamic taxonomies, which show a high level of user satisfaction. However, these studies do not address what we