Dynamic Taxonomies for Intelligent Information Access

Giovanni M. Sacco
Università di Torino, Italy

INTRODUCTION

End-user interactive access to complex information is a key requirement in most applications, from knowledge management, to e-commerce, to portals. Traditionally, only access paradigms based on the retrieval of data on the basis of precise specifications have been supported. Examples include queries on structured databases and information retrieval. There is now a growing perception that this type of paradigm does not model a large number of search tasks, such as product selection in e-commerce sites among many others, that are imprecise and require exploration, weighting of alternatives and information thinning. The recent debate on findability (Morville, 2002) and the widespread feeling that "search does not work" and "information is too hard to find" shows evidence of the crisis of traditional access paradigms.

New access paradigms supporting exploration are needed. Because the goal is end-user interactive access, a holistic approach in which modeling, interface and interaction issues are considered together, must be used and will be discussed in the following.

BACKGROUND

Four retrieval techniques are commonly used: (a) information retrieval (IR) systems (van Rijsbergen, 1979), also search engines; (b) queries on structured databases; (c) hypertext/hypermedia links and d) static taxonomies, such as Yahoo!.

IR systems exhibit an extremely wide semantic gap between the user model (concepts) and the model used by commercial retrieval systems (words). This leads to a significant loss of relevant information (Blair & Maron, 1985), and to poor user interaction because query formulation is difficult and no or very little assistance is given. In addition, because results are presented as a flat list with no systematic organization, no exploration is possible. Database queries require structured data and are not applicable to situations in which information are textual and not structured or loosely structured. Exploration is usually limited to sorting flat result lists according to different ordering criteria.

Hypermedia techniques (Groenbaek & Trigg, 1994) have become pervasive and support exploration. However, they do not support abstraction so that exploration is performed one-document-at-a-time, which is quite time consuming. Building and maintaining nontrivial hypermedia networks is very expensive.

Traditional taxonomies are based on a hierarchy of concepts that can be used to select areas of interest and restrict the portion of the infobase to be retrieved. They are easily understood by end-users, but they are not scalable for large information bases (Sacco, 2006b), so that the average number of documents retrieved becomes rapidly too large for manual inspection.

A more recent approach is the Semantic Web (Berners-Lee, Hendler, & Lassila, 2001). Although one of the driving forces behind it is retrieval, the general semantic schemata proposed are intended for programmatic access and are known to be difficult to understand and manipulate by the casual user. User interaction must be mediated by specialized agents, which increases costs, time to market and decreases the transparence and flexibility of user access.

DYNAMIC TAXONOMIES

Dynamic taxonomies (Sacco, 1987, 2000), also called *faceted search systems*, are a general knowledge management model based on a multidimensional classification of heterogeneous data items and are used to explore/browse complex information bases in a guided yet unconstrained way through a visual interface.

The intension of a dynamic taxonomy is a taxonomy designed by an expert. This taxonomy is a concept hierarchy going from the most general to the most specific concepts. A dynamic taxonomy does not require any other relationships in addition to *subsumptions* (e.g., IS-A and PART-OF relationships). Directed acyclic graph taxonomies modeling multiple inheritance are supported but rarely required.

In the extension, items can be freely classified under n (n>1) concepts at any level of abstraction (i.e., at any level in the conceptual tree). The multidimensional classification required by dynamic taxonomies is a generalization of the monodimensional classification scheme used in conventional taxonomies and models common real-life situations. First, items are very often about different concepts: for example, a news item on September 11th, 2001 can be classified under "terrorism," "airlines," "USA," and so forth. Second, items to be classified usually have different features, "perspectives"

or facets (e.g., Time, Location, etc.), each of which can be described by an independent taxonomy.

In dynamic taxonomies, a concept C is just a label that identifies all the items classified under C. Because of the subsumption relationship between a concept and its descendants, the items classified under C (items(C)) are all those items in the *deep extension* of C, that is, the set of items identified by C includes the *shallow extension* of C (i.e., all the items directly classified under C) union the deep extension of C's sons. By construction, the shallow and the deep extension for a terminal concept are the same. This set-oriented approach implies that logical operations on concepts can be performed by the corresponding set operations on their extension, and therefore the user is able to restrict the information base (and to create derived concepts) by combining concepts through all the standard logical operations (and, or, not).

A fundamental feature of this model is that dynamic taxonomies can find all the concepts related to a given concept C: these concepts represent the conceptual summary of C. Concept relationships other than subsumptions are inferred on the basis of empirical evidence through the extension only, according to the following *extensional inference rule*: two concepts A and B are related if there is at least one item *d* in the knowledge base which is classified at the same time under A or under one of A's descendants and under B or under one of B's descendants. For example, we can infer an unnamed relationship between *terrorism* and *New York*, if an item classified under *terrorism* and *New York* exists. At the same time, because *New York* is a descendant of *USA*, also a relationship between *terrorism* and *USA* can be inferred.

The extensional inference rule can be easily extended to cover the relationship between a given concept C and a concept expressed by an arbitrary subset S of the universe: C is related to S if there is at least one item d in S which is also in items(C). Hence, the extensional inference rule can produce conceptual summaries not only for base concepts, but also for any logical combination of concepts. In addition, because it is immaterial how S is produced, dynamic taxonomies can produce summaries for sets of items produced by other retrieval methods such as database queries, shape retrieval, and so forth, and therefore access through dynamic taxonomies can be easily combined with any other retrieval method.

Dynamic taxonomies are defined in terms of conceptual descriptions of items, so that heterogeneous items of any type and format can be managed in a single, coherent framework. Finally, because concept C is just a label that identifies the set of the items classified under C, concepts are language-invariant, and multilingual access can be easily supported by maintaining different language directories, holding language-specific labels for each concept in the taxonomy.

Exploration

The user is initially presented with a tree representation of the initial taxonomy for the entire knowledge base. The initial user focus F is the universe, that is, all the items in the information base. In the simplest case, the user selects a concept C in the taxonomy and zooms over it. The *zoom* operation changes the current state in the following way:

1. Concept C is used to refine the current *user focus* F, which becomes F∩items(C). Items not in the focus are discarded.
2. The tree representation of the taxonomy is modified in order to summarize the new focus. All and only the concepts related to F are retained and the count for each retained concept C' is updated to reflect the number of items in the focus F that are classified under C'. The *reduced taxonomy* is derived from the initial taxonomy by pruning all the concepts not related to F, and it is a conceptual summary of the set of documents identified by F, exactly in the same way as the original taxonomy was a conceptual summary of the universe. In fact, the term *dynamic taxonomy* indicates that the taxonomy can dynamically adapt to the subset of the universe on which the user is focusing, whereas traditional, static taxonomies can only describe the entire universe.

The retrieval process can be seen as an iterative thinning of the information base: the user selects a focus, which restricts the information base by discarding all the items not in the current focus. Only the concepts used to classify the items in the focus and their ancestors are retained. These concepts, which summarize the current focus, are those and only those concepts that can be used for further refinements. From the human computer interaction point of view, the user is effectively guided to reach his goal by a clear and consistent listing of all possible alternatives, and, in fact, this type of interaction is often called *guided thinning* or *guided navigation.* Such an iterative refinement terminates when the number of items in the focus is sufficiently small for manual inspection. In order to assist the user in deciding whether a simple concept expansion or a zoom operation is required, each concept label usually shows a count of all the items classified under it, that is, the cardinality of items(C) for all C's.

Dynamic taxonomies can be integrated with other retrieval methods in two basic ways. First, focus restrictions on the dynamic taxonomy can provide a context on which other retrieval methods can be applied, thereby increasing the precision of subsequent searches. Second, the user can start from an external retrieval method, and see a conceptual summary of the concepts that describe the result. Concepts in this summary can be used to set additional foci. These