# Dynamic Taxonomies and Intelligent User-Centric Access to Complex Portal Information

**Giovanni M. Sacco**
*Università di Torino, Italy*

## INTRODUCTION

One of the key requirements of portals is easy access to information, or *findability* according to Morville's definition (Morville, 2002). After a decade of using traditional access paradigms, such as queries on structured database systems and information retrieval or search engines, the feeling that "search does not work" and "information is too hard to find" is now reaching a consensus level. The problem is that traditional access paradigms are not suited to most search tasks, that are exploratory and imprecise in essence: the user needs to explore the information base, find relationships among concepts and think alternatives out in a guided way.

New access paradigms supporting exploration are needed. Since 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) techniques (van Rijsbergen, 1979) recently dubbed search engines; (b) queries on structured databases; (c) hypertext/hypermedia links and (d) static taxonomies, such as Yahoo!.

The limitations of IR techniques are well known: a 1985 study reported that only 20% of relevant documents were actually retrieved (Blair & Maron, 1985). Such a significant loss of information is due to the extremely wide semantic gap between the user model (concepts) and the model used by commercial retrieval systems (words). Other problems include poor user interaction because the user has to formulate his query with no or very little assistance, and no exploration capabilities since results are presented as a flat list with no systematic organization. Database queries require structured data and are not easily applicable to situations, such as portals, in which most information is textual and not structured or loosely structured.

Hypermedia (see Groenbaek & Trigg, 1994) is quite flexible, but it gives no systematic picture of relationships among documents; exploration is performed one document at a time, which is quite time consuming; and building and maintaining complex 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. Taxonomies support abstraction and are easily understood by end-users. However, they are not scalable for large information bases (Sacco, 2002), and the average number of documents retrieved becomes rapidly too large for manual inspection.

Solutions based on semantic networks, ontologies, and Semantic Web (Berners-Lee et al., 2001) are more powerful than plain taxonomies. However, general semantic schemata 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, 1998, 2000, also called *faceted classification 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. Directed acyclic graph taxonomies modeling multiple inheritance are supported but rarely required. A dynamic taxonomy does not require any other relationships in addition to *subsumptions* (e.g., IS-A and PART-OF relationships).

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). This multidimensional classification is a generalization of the mono-dimensional 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 11, 2001, can be classified under "terrorism," "airlines," "USA," and so forth. Second, items to be classified usu-

ally 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.

There are two important immediate consequences of this approach. First, since concepts identify sets of items, logical operations on concepts can be performed by the corresponding set operations on their extension. This means that the user is able to restrict the information base (and to create derived concepts) by combining concepts through the normal logical operations (and, or, not). Second, 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 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, since *New York* is a descendant of *USA*, also a relationship between *terrorism* and *USA* can be inferred. The extensional inference rule can be seen as a device to infer relationships on the basis of empirical evidence.

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. Since 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 work on conceptual descriptions of items, so that heterogeneous items of any type and format can be managed in a single, coherent framework. Finally, since 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. If the metadata descriptors used to describe an item use concepts from the taxonomy, then also the actual description of an item can be translated on the fly to different languages.

## Exploration

The user is initially presented with a tree representation of the initial taxonomy for the entire knowledge base. Each concept label has also a count of all the items classified under it, i.e., the cardinality of items(C) for all C's. The initial user focus F is the universe, i.e., all the items in the information base.

In the simplest case, the user selects a concept C in the taxonomy and *zoom* over it. The zoom operation changes the current state in two ways. First, concept C is used to refine the current *user focus* F, which becomes F∩items(C). Items not in the focus are discarded. Second, 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*.

Figures 1 to 5 show how the zoom operation works. Figure 1 shows a dynamic taxonomy: the upper half represents the intension with circles representing concepts; the lower half is the extension, and documents are represented by rectangles. Arcs going down represent subsumptions; arcs going up represent classifications. In order to compute all the concepts related to H, we first find, in Figure 2, all the documents classified under H (that is, the deep extension of H, items(H)) by following all the arcs incident to H (and, in general, its descendants): items(H) = { b, c, d }. All the items not in the deep extension of H (Figure 3) are removed from the extension. In Figure 4, the set of all the concepts