Extended Faceted Taxonomies For Web Catalogs

by Yannis Tzitzikas, Nicolas Spyratos, Panos Constantopoulos and Anastasia Analyti

What do you prefer to remember: 1000 individual terms or 3 facets of 10 terms each?

One way of designing a taxonomy for a knowledge domain is by identifying a number of different aspects, or *facets,* of the domain and then designing one taxonomy for each facet. Several studies in information, library and cognitive science have shown that in almost every knowledge domain we can indeed distinguish a number of facets, or planes of understanding. A faceted taxonomy is actually a set of taxonomies, called *facets,* where a taxonomy is a set of terms structured by a specialization/generalization relation. Using a faceted taxonomy, the indexing of objects is done by associating each object with a *compound term*, i.e. with a combination of terms coming from different facets. It has been recognized long ago that a faceted taxonomy has several advantages by comparison to a single hierarchical taxonomy, such as conceptual clarity, compactness and scalability. For example, consider two schemes for indexing the objects of a domain, the first using a single taxonomy consisting of 100 terms, and the second using a faceted taxonomy consisting of 10 facets each having 10 terms. The first scheme has 100 indexing terms while the second has 10^{10}, i.e. 10 billion, compound indexing terms! Although both schemes have the same storage requirements, i.e. each one requires storing 100 terms, the indexing terms of the second are tremendously more than the indexing terms of the first.

However, faceted taxonomies have a major drawback that prevents their deployment and use for real and large-scale applications like the Web. This drawback comes from the fact that one can form a large number of invalid compound terms, i.e. combinations of terms that do not apply to any object of the underlying domain. For example, refer to the faceted taxonomy for a tourist information application shown in Figure 1 and consider the terms WinterSports from the facet *Sports* and Crete from the facet *Location.* The compound term WinterSports.Crete is *invalid* as there is never enough snow in Crete! In contrast, the compound term SeaSports.Crete is certainly valid. The inability to infer the valid compound terms may create problems in object indexing (laborious and/or erroneous indexing), and in browsing (an invalid compound term will yield no objects). Due to such problems, existing Web catalogs have a strictly hierarchical structure like the one of Figure 2. Such taxonomies suffer from several problems such as incomplete terminology, huge size (for example the taxonomy of Open Directory consists of 300.000 terms), and confusing structure.

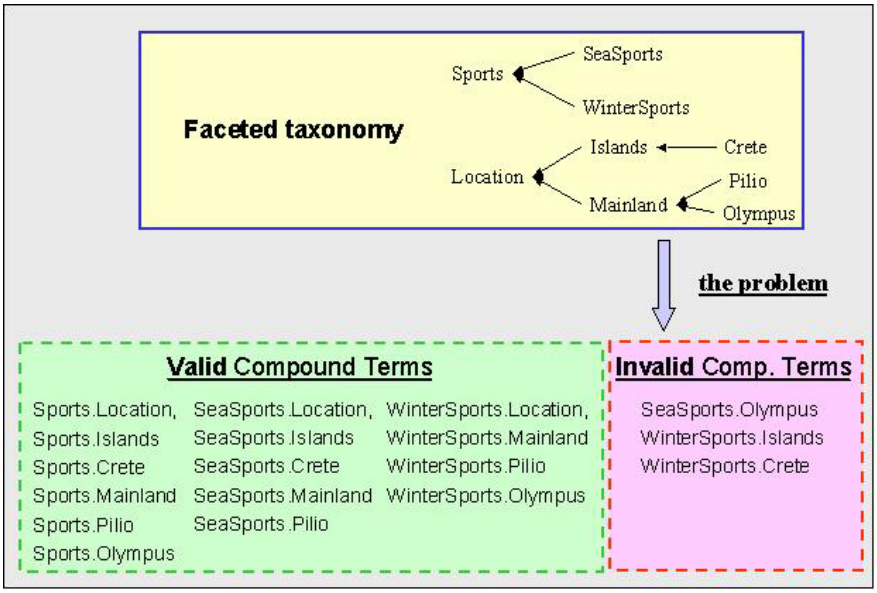

Figure 1 A faceted taxonomy consisting of two facets, *Sports* and *Location*, and the partition of the set of compound terms to the set of valid and the set of invalid terms.

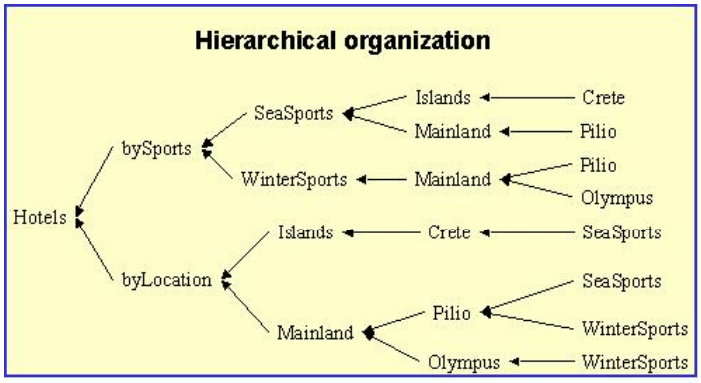

Figure 2 A hierarchical organization of the valid compound terms of the faceted taxonomy of Figure 1.

Being able to infer the validity of a compound term in a faceted taxonomy would be very useful for facilitating the indexing and for preventing errors, especially in cases where indexing is done by many human editors (indexers). For example, Web pages in the Open Directory are currently indexed by more than 20.000 volunteers. Moreover, if we could infer the valid compound terms in a faceted taxonomy then we would be able to generate navigation trees (like the taxonomies of the Web catalogs) on the *fly*, consisting of nodes that correspond to valid compound terms. However, defining manually the set of valid compound terms even for facets of relatively small size, would be a formidable task for the designer. For example, Figure 1 shows the partition of compound terms into sets of valid and invalid terms.

To alleviate this problem, we have defined two extensions of faceted taxonomies, which we call *PEFT* and *NEFT*, which allow specifying the valid compound terms in a flexible and efficient manner. A *PEFT* (Positive Extended Faceted Taxonomy) is a faceted taxonomy enriched with a set *P* of certainly valid compound terms, while a *NEFT* (Negative Extended Faceted Taxonomy) is a faceted taxonomy enriched with a set *N* of certainly invalid compound terms. The designer simply declares a small set of valid or invalid compound terms and other (valid or invalid) compound terms are then inferred by an inference mechanism based on semantic implication. Figure 3 shows how we can specify the valid compound terms of the faceted taxonomy of Figure 1, i.e. the sets "Valid Compound Terms " and "Invalid Compound Terms" as enumerated in that figure, by employing either a *PEFT,* or a *NEFT*. In each case we can derive *dynamically* a navigation tree such as the one shown in Figure 2 which can be exploited during object indexing and browsing.

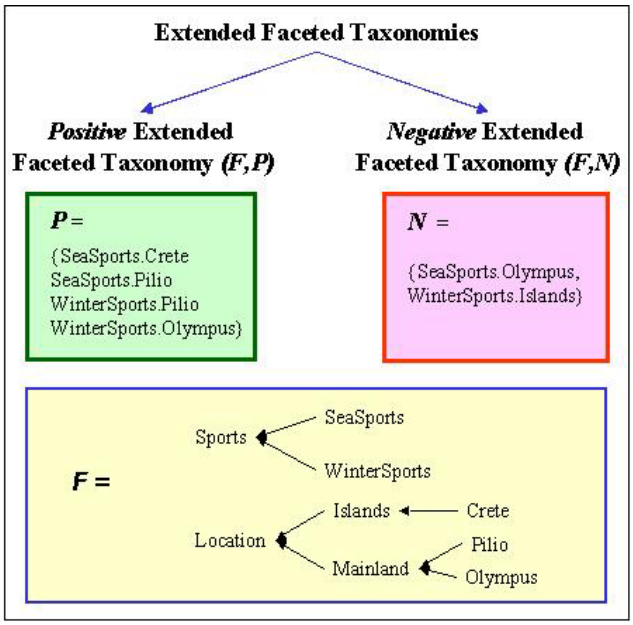

Figure 3 Two extended faceted taxonomies

Our approach can be used for developing Web catalogs that offer complete navigation trees, require less storage space, and are more comprehensive and scalable. Furthermore, taxonomies designed according to our approach can be integrated or articulated more easily than the hierarchical ones. Further research includes extending this approach so as to define an algebra over taxonomies with operators that allow the specification of the valid compound terms, using both positive and negative sets of compound terms.

This work is result of basic research conducted in ICS-FORTH the last two years.

Please contact:
Yannis Tzitzikas, Information Systems Laboratory, ICS-FORTH
Tel: +30 810 391 623
E-mail: tzitzik@csi.forth.gr
Web page: http://www.csi.forth.gr/~tzitzik