23rd International Conference. on Conceptual Modeling, ER'2004, Shanghai, China, November 2004.

FASTAXON: A System for FAST (and Faceted) TAXONomy Design

Yannis Tzitzikas[1], Raimo Launonen[1], Mika Hakkarainen[1], Pekka Korhonen[2], Tero Leppänen[2], Esko Simpanen[2], Hannu Törnroos[2], Pekka Uusitalo[2], Pentti Vänskä[2]

[1] *VTT Information Technology, P.O.Box 1201, 02044 VTT, Finland*
[2] *Helsinki University of Technology, Finland*
Contact emails: ytz@info.fundp.ac.be, Raimo.Launonen@vtt.fi, Mika.Hakkarainen@vtt.fi

Building very big taxonomies is a laborious task vulnerable to errors and management/scalability deficiencies. FASTAXON is a system for building very big taxonomies in a quick, flexible and scalable manner that is based on the *faceted* classification paradigm [4] and the *Compound Term Composition Algebra* [5]. Below we sketch the architecture and the functioning of this system and we report our experiences from using this system in real applications.

Taxonomies, i.e. hierarchies of names, is probably the oldest and most widely used conceptual modeling tool still used in Web directories, Libraries and the Semantic Web (e.g. see XFML [1]). Moreover, the advantages of the taxonomy-based conceptual modeling approach for building large scale *mediators* and *P2P systems* that support semantic-based retrieval services have been analyzed and reported in [7, 6, 8]. However, building very big taxonomies is a laborious task vulnerable to errors and management/scalability deficiencies. One method for building efficiently a very big taxonomy is to first define a *faceted taxonomy* (i.e. a set of independently defined taxonomies called *facets*) like the one presented in Figure 1, and then derive automatically the inferred *compound taxonomy* i.e. the taxonomy of all possible *compound terms* (conjunctions of terms) over the faceted taxonomy. Faceted taxonomies carry a number of well known advantages over single hierarchies in terms of building and maintaining them, as well as using them in multicriteria indexing (e.g. see [3]). FASTAXON is a system for building big (compound) taxonomies based on the above mentioned idea. Using the system, the designer at first defines a number of facets and assigns to each one of them one taxonomy. After that the system can generate dynamically (and on the fly) a navigation tree that allows to the designer (as well to the object indexer or end user) to browse the set of *all* possible compound terms.

A drawback, however, of faceted taxonomies is the cost of avoiding the *invalid* (meaningless) compound terms, i.e. those that do not apply to any object in the domain. Let's consider the faceted taxonomy of Figure 1. Clearly we cannot do any winter sport in the Greek islands (Crete and Cefalonia) as they never have enough snow, and we cannot do any sea sport in Olympus because Olympus is a mountain. For the sake of this example, let us also suppose that only

Cefalonia has a Casino. According to this assumption, the partition of the set of compound terms to the set of *valid* (meaningful) and *invalid* (meaningless) is shown in Table 1. The availability of such a partition would be very useful during the construction of a materialized faceted taxonomy (i.e. a catalog based on a faceted taxonomy). It could be exploited in the indexing process for *preventing* indexing errors, i.e. for allowing only meaningful compound terms to be assigned to objects. It could also *aid* the indexer during the indexing process, by generating dynamically a single hierarchical navigation tree that allows selecting the desired compound term by browsing *only* the meaningful compound terms. However, even from this toy example, it is more than obvious that the definition of such a partition would be a formidably laborious task for the designer.

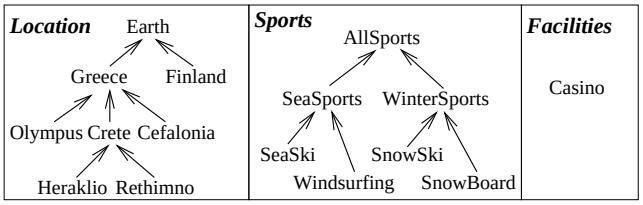

Fig. 1. A faceted taxonomy for indexing hotel Web pages

FASTAXON allows specifying the meaningful compound terms in a very flexible manner. It is the first system that implements the recently emerged *Compound Term Composition Algebra* (CTCA) [5]. This allows to the designer to use an algebraic expression for specifying the valid compound terms. This involves declaring only a *small* set of valid or invalid compound terms from which other (valid or invalid) compound terms are then *inferred*. For instance, the partition shown in Table 1, can be defined using the expression:
$e = (Location \ominus_N Sports) \oplus_P Facilities$ with the following P and N parameters:
$N = \{\{Crete, WinterSports\}, \{Cefalonia, WinterSports\}\}$,
$P = \{\{Cefalonia, SeaSki, Casino\}, \{Cefalonia, Windsurfing, Casino\}\}$.
Specifically, FASTAXON provides an Expression Builder for formulating CTCA expressions in a flexible, interactive and guided way. Only the expression that defines the desired compound terminology is stored (and not the inferred partition), as an inference mechanism is used to check (in polynomial time) whether a compound term belongs to the compound terminology of the expression.

The productivity obtained using FASTAXON is quite impressive. The so far experimental evaluation has shown that in many cases a designer can define from scratch a compound taxonomy of around 1000 indexing terms in some minutes. FASTAXON has been implemented as a client/server Web-based system written in Java. The server is based on the Apache Web server, the Tomcat application server and uses MySQL for persistent storage. The user interface is based on DHTML (dynamic HTML), JSP (Java Server Pages) and Java Servlet technologies (J2EE). The client only needs a Web browser that support JavaScripts (e.g. Microsoft Internet Explorer 6). Future extensions include modules for importing and exporting XFML [1] and XFML+CAMEL [2] files. FASTAXON will

Valid	
Earth, AllSports	Greece, AllSports
Finland, AllSports	Olympus, AllSports
Crete, AllSports	Cefal., AllSports
Reth., AllSports	Heraklio, AllSports
Earth, SeaSports	Greece, SeaSports
Finland, SeaSports	Crete, SeaSports
Cefal., SeaSports	Reth., SeaSports
Heraklio, SeaSports	Earth, WinterSp.
Greece, WinterSp.	Finland, WinterSp.
Olympus, WinterSp.	Earth, SeaSki
Greece, SeaSki	Finland, SeaSki
Crete, SeaSki	Cefal., SeaSki
Reth., SeaSki	Heraklio, SeaSki
Earth, WindSurf.	Greece, WindSurf.
Finland, WindSurf.	Crete, WindSurf.
Cefal., WindSurf.	Reth., WindSurf.
Heraklio, WindSurf.	Earth, SnowB.
Greece, SnowB.	Finland, SnowB.
Olympus, SnowB.	Earth, SnowSki
Greece, SnowSki	Finland, SnowSki
Olympus, SnowSki	Earth, AllSports, Cas.
Greece, AllSports, Cas.	Cefal., AllSports, Cas.
SeaSports, Cas.	SeaSports, Cas.
Cefal., SeaSports, Cas.	Earth, WinterSp., Cas.
Greece, WinterSp., Cas.	Earth, SeaSki, Cas.
Greece, SeaSki, Cas.	Cefal., SeaSki, Cas.
Earth, WindSurf., Cas.	Greece, WindSurf., Cas.
Cefal., WindSurf., Cas.	Earth, SnowB., Cas.
Greece, SnowB., Cas.	Earth, SnowSki, Cas.
Greece, SnowSki, Cas.	

Invalid	
Olympus, SeaSports	
Crete, WinterSp.	Cefal., WinterSp.
Reth., WinterSp.	Heraklio, WinterSp.
Olympus, SeaSki	Olympus, WindSurf.
Crete, SnowB.	Cefal., SnowB.
Reth., SnowB.	Heraklio, SnowB.
Crete, SnowSki	Cefal., SnowSki
Reth., SnowSki	Heraklio, SnowSki
Olympus, SeaSports, Cas.	Crete, WinterSp., Cas.
Cefal., WinterSp., Cas.	Reth., WinterSp., Cas.
Heraklio, WinterSp., Cas.	Olympus, SeaSki, Cas.
Olympus, WindSurf., Cas.	Crete, SnowB., Cas.
Cefal., SnowB., Cas.	Reth., SnowB., Cas.
Heraklio, SnowB., Cas.	Crete, SnowSki, Cas.
Cefal., SnowSki, Cas.	Reth., SnowSki, Cas.
Heraklio, SnowSki, Cas.	Olympus, AllSports, Cas.
Crete, AllSports, Cas.	Reth., AllSports, Cas.
Heraklio, AllSports, Cas.	Crete, SeaSports, Cas.
Reth., SeaSports, Cas.	Heraklio, SeaSports, Cas.
Olympus, WinterSp., Cas.	Crete, SeaSki, Cas.
Reth., SeaSki, Cas.	Heraklio, SeaSki, Cas.
Crete, WindSurf., Cas.	Reth., WindSurf., Cas.
Heraklio, WindSurf., Cas.	Olympus, SnowB., Cas.
Olympus, SnowSki, Cas.	Finland, AllSports, Cas.
Finland, SeaSports, Cas.	Finland, WinterSp., Cas.
Finland, SeaSki, Cas.	Finland, WindSurf., Cas.
Finland, SnowSki, Cas.	Finland, SnowB., Cas.

Table 1. The Valid and Invalid compound terms of the example of Figure 1

be published under the VTT Open Source Licence within 2004 (for more see http://fastaxon.erve.vtt.fi/).

References

1. "XFML: eXchangeable Faceted Metadata Language". http://www.xfml.org.
2. "XFML+CAMEL:Compound term composition Algebraically-Motivated Expression Language". http://www.csi.forth.gr/markup/xfml+camel.
3. Ruben Prieto-Diaz. "Implementing Faceted Classification for Software Reuse". *Communications of the ACM*, 34(5):88–97, 1991.
4. S. R. Ranganathan. "The Colon Classification". In Susan Artandi, editor, *Vol IV of the Rutgers Series on Systems for the Intellectual Organization of Information*. New Brunswick, NJ: Graduate School of Library Science, Rutgers University, 1965.
5. Y. Tzitzikas, A. Analyti, N. Spyratos, and P. Constantopoulos. "An Algebraic Approach for Specifying Compound Terms in Faceted Taxonomies". In *Information Modelling and Knowledge Bases XV, Procs of EJC'03*, pages 67–87. IOS Press, 2004.
6. Y. Tzitzikas and C. Meghini. "Ostensive Automatic Schema Mapping for Taxonomy-based Peer-to-Peer Systems". In *7th Int. Workshop on Cooperative Information Agents, CIA-2003*, pages 78–92, Helsinki, Finland, August 2003.
7. Y. Tzitzikas, C. Meghini, and N. Spyratos. "Taxonomy-based Conceptual Modeling for Peer-to-Peer Networks". In *Procs of 22th Int. Conf. on Conceptual Modeling, ER'2003*, pages 446–460, Chicago, Illinois, October 2003.

8. Y. Tzitzikas, N. Spyratos, and P. Constantopoulos. "Mediators over Taxonomy-based Information Sources". *VLDB Journal*, 2004. (to appear).