

# Faceted Taxonomy-based Information Management

Yannis Tzitzikas<sup>1,2</sup> and Anastasia Analyti<sup>2</sup>

<sup>1</sup> Department of Computer Science, University of Crete, Greece, and

<sup>2</sup> Institute of Computer Science, FORTH-ICS, Crete, Greece

{tzitzik, analyti}@ics.forth.gr

## Abstract

Faceted indexing and searching are being increasingly studied in the literature and used for real-life applications, e.g., for publishing heterogeneous museum collections on the Web. In this paper, we discuss in brief several aspects of managing (faceted) taxonomy-based information sources. Specifically, we discuss (i) the semantic description of faceted taxonomies, based on the Compound Term Composition Algebra (CTCA), (ii) the revision of CTCA expressions, as faceted taxonomies evolve, (iii) the dynamic generation of navigational trees (and other applications of CTCA), and (iv) the integration and personalization of taxonomy-based sources.

## 1. Introduction

A *faceted taxonomy* is a set of taxonomies, each one describing the domain of interest from a different (preferably orthogonal) point of view [6]. Having a faceted taxonomy, each domain object (e.g., a book or a Web page) can be indexed using a *compound term*, i.e., a set of terms from the different facets. A *materialized faceted taxonomy* (or *faceted taxonomy-based source*) is a faceted taxonomy accompanied by a set of object indexes. Figure 1 shows an indicative materialized faceted taxonomy that consists of three facets and indexes two hotel Web pages.

Faceted taxonomies are used in Marketplaces [9], Libraries, Software Repositories [5], e-government portals [8], publishing museum collections on the Semantic Web [3], and several other application domains. Current interest in faceted taxonomies is also indicated by several recent projects (like FATKS<sup>1</sup>, FACET<sup>2</sup>, FLAMENGO<sup>3</sup>, SemWeb<sup>4</sup>, SWED<sup>5</sup>) and the emergence of XFML [1] (Core-

eXchangeable Faceted Metadata Language), a markup language for applying the faceted classification paradigm on the Web.

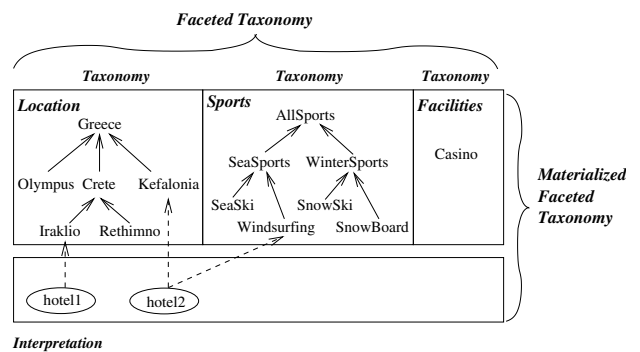


Figure 1. A materialized faceted taxonomy

In this paper, we present an overview of our recent research on faceted taxonomy-based information management. In particular, in Section 2, we describe the *Compound Term Composition Algebra* (CTCA), an algebra that allows specifying the set of meaningful compound terms (i.e., meaningful conjunctions of terms) over a faceted taxonomy in a flexible and efficient manner. In Section 3, we discuss the dynamic construction of user navigational trees, based on CTCA expressions, the automatic derivation of the shortest CTCA expression that describes a materialized faceted taxonomy, and several other applications of CTCA. In Section 4, we describe how CTCA expressions should be revised in the case that the faceted taxonomy is updated. Subsequently, in Section 5, we discuss the issue of automatically articulating or linking taxonomies based on their common instances. The resulting mappings can alleviate the cost of defining inter-taxonomy mappings in taxonomy-based mediator and P2P systems. Query evaluation methods for these kinds of systems are also investigated. Additionally, we discuss the personalization of taxonomy-based sources. Finally, in Section 6, we conclude the paper and sketch directions that deserve further research.

<sup>1</sup><http://www.ucl.ac.uk/fatks/database.htm>

<sup>2</sup>[http://www.glam.ac.uk/soc/research/hypermedia/facet\\_proj/index.php](http://www.glam.ac.uk/soc/research/hypermedia/facet_proj/index.php)

<sup>3</sup><http://bailando.sims.berkeley.edu/flamenco.html>

<sup>4</sup><http://www.seco.tkk.fi/projects/semweb/>

<sup>5</sup><http://www.swed.org.uk/>

Name	Notation	Definition
terminology	$\mathcal{T}$	a set of names, called <i>terms</i>
subsumption	$\leq$	a preorder relation (reflexive and transitive)
taxonomy	$(\mathcal{T}, \leq)$	$\mathcal{T}$ is a terminology, $\leq$ a subsumption relation over $\mathcal{T}$
faceted taxonomy	$\mathcal{F} = \{F_1, \dots, F_k\}$	$F_i = (\mathcal{T}_i, \leq_i)$ , for $i = 1, \dots, k$ and all $\mathcal{T}_i$ are disjoint
compound term over $\mathcal{T}$	$s$	any subset of $\mathcal{T}$ (i.e., any element of $\mathcal{P}(\mathcal{T})$ )
compound terminology	$S$	a subset of $\mathcal{P}(\mathcal{T})$ that includes $\emptyset$
compound ordering	$\preceq$	$s \preceq s'$ iff $\forall t' \in s' \exists t \in s$ such that $t \leq t'$
broaders of $s$	$\text{Br}(s)$	$\{s' \in \mathcal{P}(\mathcal{T}) \mid s \preceq s'\}$
narrowers of $s$	$\text{Nr}(s)$	$\{s' \in \mathcal{P}(\mathcal{T}) \mid s' \preceq s\}$
broaders of $S$	$\text{Br}(S)$	$\cup \{\text{Br}(s) \mid s \in S\}$
narrowers of $S$	$\text{Nr}(S)$	$\cup \{\text{Nr}(s) \mid s \in S\}$
object domain	$\text{Obj}$	any denumerable set of objects
interpretation of $\mathcal{T}$	$I$	any function $I : \mathcal{T} \rightarrow 2^{\text{Obj}}$
model of $(\mathcal{T}, \leq)$ induced by $I$	$\bar{I}$	$\bar{I}(t) = \cup \{I(t') \mid t' \leq t\}$
materialized faceted taxonomy	$(\mathcal{F}, I)$	$\mathcal{F}$ is a faceted taxonomy $\{F_1, \dots, F_k\}$ , $I$ is an interpretation of $\mathcal{T} = \bigcup_{i=1,k} \mathcal{T}_i$

**Table 1. Basic notions and notations**

## 2 Compound Term Composition Algebra

The *Compound Term Composition Algebra (CTCA)* was proposed for defining the meaningful compound terms over a faceted taxonomy in a flexible and efficient manner. The problem of meaningless compound terms and the effort needed to specify the meaningful ones is a practical problem identified even by Ranganathan himself [6] (about 80 years ago). This is probably the main reason why faceted taxonomies have not dominated every application domain despite their uncontested advantages over the single taxonomies. CTCA is the only well-founded and flexible solution to this problem. Table 1 recalls in brief the basic notions around taxonomies, faceted taxonomies, and materialized faceted taxonomies (for more refer to [14, 13]).

CTCA has four basic algebraic operations, namely, *plus-product* ( $\oplus$ ), *minus-product* ( $\ominus$ ), *plus-self-product*, ( $\oplus^*$ ), and *minus-self-product* ( $\ominus^*$ ). All these are operations over  $\mathcal{P}(\mathcal{T})$ , the powerset of  $\mathcal{T}$ , where  $\mathcal{T}$  is the union of the terminologies of all facets. The initial operands, thus the building blocks, of the algebra are the *basic compound terminologies*, defined as:  $T_i = \text{Br}(\{\{t\} \mid t \in \mathcal{T}_i\})$ , where  $\mathcal{T}_i$  is the terminology of the facet  $F_i$ ,  $i = 1, \dots, k$ . An expression  $e$  over  $\mathcal{F}$  is defined according to the following grammar:

$$e ::= \oplus_P(e, \dots, e) \mid \ominus_N(e, \dots, e) \mid \oplus_P^* T_i \mid \ominus_N^* T_i \mid T_i,$$

where the parameters  $P$  and  $N$  denote sets of valid and invalid compound terms over the range of the operation, respectively. Roughly, CTCA allows specifying the valid compound terms over a faceted taxonomy by providing a small set of valid (parameter  $P$ ) and a small set of invalid (parameter  $N$ ) compound terms. The self-product operations allow specifying the meaningful compound terms over one facet. Specifically, the definition of each operation of

CTCA is summarized in Table 2, where  $S_i$ ,  $i = 1, \dots, n$ , are compound terminologies. If  $e$  is an expression,  $S_e$  denotes the outcome of this expression and is called the *compound terminology* of  $e$ . In addition,  $(S_e, \preceq)$  is called the *compound taxonomy* of  $e$ .

An expression  $e$  is *well formed* iff every facet appears at most once in  $e$ , and the parameter sets  $P$  and  $N$  are always subsets of the corresponding set of *genuine compound terms*. Specifically, each parameter  $P$  (resp.  $N$ ) of an operation  $\oplus_P(e_1, \dots, e_k)$  (resp.  $\ominus_N(e_1, \dots, e_k)$ ) should be subset of the set of genuine compound terms over the compound terminologies  $S_{e_1}, \dots, S_{e_k}$ , i.e., subset of:

$$G_{S_{e_1}, \dots, S_{e_k}} = S_{e_1} \oplus \dots \oplus S_{e_k} - \cup_{i=1}^k S_{e_i}$$

From an application point of view, another important remark is that there is no need to store the set of valid compound terms that are defined by an expression, as the algorithm *IsValid*( $e, s$ ) (given in [14]) can check whether a compound term  $s$  belongs to the set of compound terms defined by an expression  $e$  (i.e., whether  $s \in S_e$ ) in polynomial time. Specifically, the computational complexity of this algorithm is  $O(|\mathcal{T}|^3 * |s| * |\mathcal{P} \cup \mathcal{N}|)$ , where  $\mathcal{P}$  denotes the union of all  $P$  parameters and  $\mathcal{N}$  denotes the union of all  $N$  parameters appearing in  $e$ <sup>6</sup>. Thus, only the faceted taxonomy  $\mathcal{F}$  and the CTCA expression  $e$  need to be stored.

As an example, recall the faceted taxonomy of Figure 1. One can easily see that several compound terms over this faceted taxonomy are meaningless, in the sense that they cannot be applied to any object of the domain. For instance, we cannot do any winter sport in the Greek islands (Crete and Kefalonia) 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 in Kefalonia there exists a hotel that has a casino,

<sup>6</sup>Note that  $|\mathcal{T}|$  is not expected to be very large, in a faceted taxonomy.

Operation	$e$	$S_e$
product	$S_1 \oplus \dots \oplus S_n$	$\{s_1 \cup \dots \cup s_n \mid s_i \in S_i\}$
plus-product	$\oplus_P(S_1, \dots, S_n)$	$S_1 \cup \dots \cup S_n \cup Br(P)$
minus-product	$\ominus_N(S_1, \dots, S_n)$	$S_1 \oplus \dots \oplus S_n - Nr(N)$
self-product	$\overset{*}{\oplus}(T_i)$	$P(T_i)$
plus-self-product	$\oplus_P(T_i)$	$T_i \cup Br(P)$
minus-self-product	$\overset{*}{\ominus}_N(T_i)$	$\overset{*}{\oplus}(T_i) - Nr(N)$

**Table 2. The operations of CTCA**

and that this hotel also offers sea ski and windsurfing sports. According to this assumption, the partition of compound terms to the set of *valid* (meaningful) compound terms and *invalid* (meaningless) compound terms can be defined using the subsequent CTCA expression:

$$e = (Location \ominus_N Sports) \oplus_P Facilities,$$

with the following  $P$  and  $N$  parameters:

$$N = \{\{Crete, WinterSports\}, \{Kefalonia, WinterSports\}\}$$

$$P = \{\{Kefalonia, SeaSki, Casino\}, \\ \{Kefalonia, Windsurfing, Casino\}\}$$

### 3 Applications of CTCA

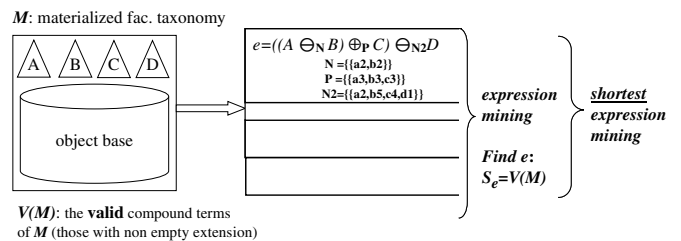
As we can infer the valid compound terms of a faceted taxonomy dynamically (through the algorithm  $IsValid(e, s)$ ), we are able to generate a single hierarchical navigation tree *on the fly*, having only valid compound terms as nodes. Intuitively, a *navigation tree* is a tree whose nodes  $n$  correspond to valid compound terms  $s$  (in the sense that both  $n, s$  index the same objects). Moreover, the navigation tree contains nodes that enable the user to start browsing in one facet and then cross to another, and so on, until reaching the desired level of specificity. The algorithm for deriving navigation trees on the fly is given in [14] and is implemented in the FASTAXON system [15]. Alternatively, we can design a user interface that consists of one subwindow per facet, and guides the user through only meaningful compound term selections. Initially, a facet subwindow lists the terms of the facet at the first-level of the facet hierarchy (considering that the zero-level is the top node of the facet). The user may select a term (*selected term*) of a facet subwindow (*selected facet*). Then, all terms that do not combine with the current selection are eliminated from the remaining facet subwindows. Additionally, the list of terms in the selected-facet subwindow is replaced by the list of children of the selected term. Previous actions can now be repeated (building step-by-step a *selected compound term*), until the user specifies the valid compound term of his/her interest. Such a user interface is presented in [19, 3]. Both of these interfaces can be used for (i) object indexing - preventing indexing errors, (ii) browsing - guiding the user to only meaningful selections, and (iii) testing

whether the derived compound taxonomy contains only the desired compound terms.

The algebra can also be used for *query optimization* in the case that object retrieval is achieved through a query language [7]. For example, consider the faceted taxonomy of Figure 1, and assume that the user wants to retrieve all hotels located in Greece and offer winter sports. As  $\{Crete, WinterSports\}$  is an invalid compound term, the system (optimizing execution) does not have to look for hotels located in Crete at all.

Another application of the algebra is *configuration management*. Consider a product whose configuration is determined by a number of parameters, each associated with a finite number of values. However, some configurations may be unsupported, unviable, or unsafe. For this purpose, the product designer can employ an expression which specifies all valid configurations, thus ensuring that the user selects only among these.

Assuming a materialized faceted taxonomy  $M$ , the problem of automatically deriving an expression  $e$  (or the shortest expression  $e$ ) that specifies all extensionally valid compound terms of  $M$ ,  $V(M)$ , is elaborated in [13]. This problem is called *expression mining* and is illustrated in Figure 2.



**Figure 2. CTCA expression mining**

This means that CTCA can be exploited both forthrightly *and* reversely, i.e., a designer can formulate an expression in order to specify quickly the set of valid compound terms, while from an existing set of valid compound terms an algorithm can find an expression that describes these compound terms. Figure 3 illustrates both scenarios. The latter direction has several other applications. For example, it can be used for reorganizing single-hierarchical taxonomies on the Web (Figure 3), for compressing large symbolic data tables (as shown in [11]), and for exchanging in a compact way the extensionally valid compound terms of a materialized faceted taxonomy.

### 4 Taxonomy Evolution and CTCA

Taxonomy updates (addition and deletion of terms or subsumption relationships) may turn a CTCA expression  $e$  ill-formed and the compound terms specified by  $e$  to no

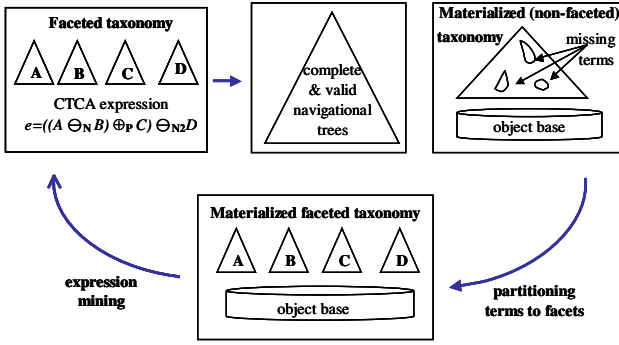


Figure 3. Various scenarios involving CTCA

longer reflect the domain knowledge originally expressed in  $e$ . In [12], we describe how we can revise a CTCA expression  $e$  after a taxonomy update, such that the new expression  $e'$  is well-formed and its semantics (defined valid compound terms) is as close as possible to the semantics of the original expression  $e$  before the update. Figure 4 illustrates the problem.

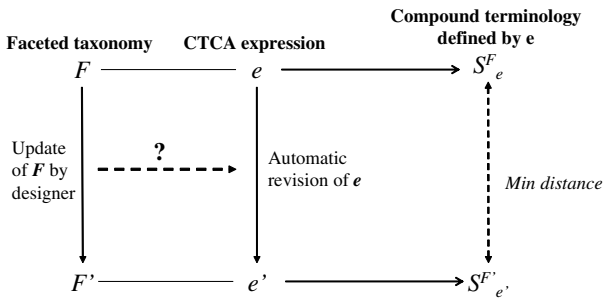


Figure 4. Revising CTCA expressions after taxonomy updates

The deletion/addition of terms or subsumption relationships in a faceted taxonomy can be handled by extending the  $P/N$  parameters of the CTCA expression  $e$ , such that missing compound terms are recovered and extra ones are removed. However, the addition of a subsumption relationship in a faceted taxonomy cannot be handled, so straightforwardly. The reason is that, since the semantics of the operations  $\oplus_P/\ominus_N$  are defined on the basis of the transitive relation  $\preceq$ , after the addition of a subsumption relationship we may no longer be able to separate (from the semantics) compound terms that were previously separable (i.e., compound terms which were not  $\preceq$ -related before the addition of the subsumption link). In such cases, the resulting compound terminology of any revised expression may neither be subset nor superset of the original compound terminology. The treatment of such cases is described in detail in [12].

## 5 Integration and Personalization of Taxonomy-based Sources

Assume that we have two taxonomy-based sources and that we want to establish mappings between their taxonomies. If they share instances then the ostensive method described in [16] can be used. Cornerstone of this method is what is called *naming function* (also analyzed in [17]). Let  $S$  be the extension of  $\bar{I}$  (for the definition of  $\bar{I}$ , see Table 1) over the set of positive queries  $Q^+$  (i.e., the set of boolean expressions of terms with no negation). As  $S$  is not always an onto function (if we consider it as a function from  $Q^+$  to the powerset of  $Obj$ ), “approximate” naming functions to a set of objects  $A$  are introduced, specifically a *lower* naming function  $n^-$  and an *upper* naming function  $n^+$ , defined as follows:

$$n^-(A) = lub\{q \in Q^+ \mid S(q) \subseteq A\}$$

$$n^+(A) = glb\{q \in Q^+ \mid S(q) \supseteq A\},$$

where  $A \subseteq Obj$ ,  $lub$  stands for least upper bound, and  $glb$  stands for greatest lower bound with respect to the query containment ordering. It is proved that:

$$n^+(A) \sim \bigvee \{D_I(o) \mid o \in A\}$$

$$n^-(A) \sim \bigvee \{D_I(o) \mid o \in A, S(D_I(o)) \subseteq A\},$$

where  $D_I(o) = \bigwedge \{t \in T \mid o \in I(t)\}$ . The time complexity for computing these names is polynomial. The ostensive method can also be used as a protocol for establishing mappings between sources that are stored distributed. In brief, in order to map a term or query  $q$  of a source  $S_1$  to a term or query of a source  $S_2$ ,  $S_1$  sends its answer  $O_1 = S(q)$  to source  $S_2$ ,  $S_2$  then computes the upper and lower name of the objects received (i.e.,  $n^-(O_1)$  and  $n^+(O_1)$ ) and sends these names (accompanied by their answers) to  $S_1$ . The latter, by comparing the object set sent with the answers received, establishes relationships between  $q$  and the received names.

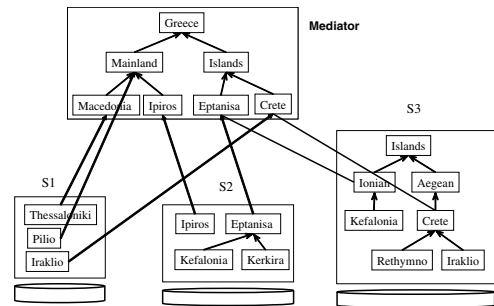
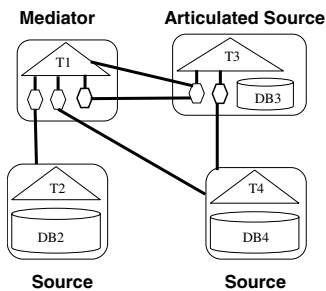


Figure 5. Taxonomy-based mediators

Suppose that we have different taxonomy-based sources and that we wish to provide a unified browsing or query interface to their indexed objects, either through one of the existing taxonomies or a new one. Then, based on inter-taxonomy mappings (defined either manually or using the ostensive method described previously), we can build a mediator system as illustrated in Figure 5. In [18], we describe



**Figure 6. Taxonomy-based P2P systems**

in detail all involved issues and provide the query evaluation algorithms for several mediator operation modes. In [4], we investigate the same problem in a peer-to-peer setting, comprising of primary sources, mediators, and articulated sources, as illustrated in Figure 6.

Concerning updates and personalization of taxonomy-based sources, [17] elaborates on the problem of updating the index of an object after user feedback. The complexity of this task is again polynomial.

## 6. Concluding Remarks and Further Research

In this paper, we discussed in brief our research on managing (faceted) taxonomy-based information sources. The Compound Term Composition Algebra (CTCA) and its applications refer mainly to faceted taxonomy-based sources. However, our work on integration and personalization can be applied equally well to both single and faceted taxonomy-based sources. Finally, we would like to mention that though we make the assumption that faceted taxonomies pre-exist, they could also be derived from statistically analyzing text corpora, as it is the case in [2, 10].

Our future plans include the extension of CTCA and its applications to a more realistic framework, where facets are not independent but interrelated through subsumption relationships between their terms. Further, we intend to investigate the problem of deriving the shortest CTCA expression that describes a faceted taxonomy, whose facet terms are defined as concepts of an OWL ontology.

## References

- [1] “XFML: eXchangeable Faceted Metadata Language”. <http://www.xfml.org>.
- [2] W. Dakka, P. G. Ipeirotis, and K. R. Wood. “Automatic Construction of Multifaceted Browsing Interfaces”. *Procs. of the 14th ACM Intern. Conf. on Information and Knowledge Management*, pages 768–775, 2005.
- [3] E. Hyvönen, E. Mäkelä, M. Salminen, A. Valo, K. Viljanen, S. Saarela, M. Junnila, and S. Kettula. “MUSEUMFINLAND - Finnish Museums on the Semantic Web”. *Journal of Web Semantics*, 3(2-3):224–241, 2005.
- [4] C. Meghini and Y. Tzitzikas. “Querying Articulated Sources”. In *Procs. of the 3rd Intern. Conf. on Ontologies, Databases and Applications of Semantics for Large Scale Information Systems (ODBASE’2004)*, pages 945–962, 2004.
- [5] R. Prieto-Diaz. “Implementing Faceted Classification for Software Reuse”. *Communications of the ACM*, 34(5):88–97, 1991.
- [6] S. R. Ranganathan. “The Colon Classification”. In S. Arntandi, 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.
- [7] K. A. Ross and A. Janevski. “Querying Faceted Databases”. In *2nd Intern. Workshop on Semantic Web and Databases, SWDB’2004 (satellite of VLDB’04)*, pages 199–218, 2004.
- [8] G. M. Sacco. “Guided Interactive Information Access for E-Citizens”. In *Procs. of the 4th Intern. Conf. on Electronic Government (EGOV-2005)*, pages 261–268, 2005.
- [9] G. M. Sacco. “The Intelligent e-Store: Easy Interactive Product Selection and Comparison”. In *Procs. of the 7th IEEE Intern. Conf. on E-Commerce Technology (CEC-2005)*, pages 240–248, 2005.
- [10] E. Stoica, M. A. Hearst, and M. Richardson. “Automating Creation of Hierarchical Faceted Metadata Structures”. In *Procs. of the Human Language Technology Conference (NAACL HLT 2007)*, 2007.
- [11] Y. Tzitzikas. “An Algebraic Method for Compressing Symbolic Data Tables”. *Journal of Intelligent Data Analysis (IDA)*, 10(4), September 2006.
- [12] Y. Tzitzikas. “Evolution of Faceted Taxonomies and CTCA Expressions”. *Knowledge and Information Systems (KAIS)*, 2006. (accepted for publication).
- [13] Y. Tzitzikas and A. Analyti. “Mining the Meaningful Term Conjunctions from Materialised Faceted Taxonomies: Algorithms and Complexity”. *Knowledge and Information Systems Journal (KAIS)*, 9(4):430–467, May 2006.
- [14] Y. Tzitzikas, A. Analyti, N. Spyrtatos, and P. Constantopoulos. “An algebra for specifying valid compound terms in faceted taxonomies”. *Data & Knowledge Engineering*, 62(1):1–40, 2007.
- [15] Y. Tzitzikas, R. Launonen, M. Hakkarainen, P. Kohonen, T. Leppanen, E. Simpanen, H. Tornroos, P. Uusitalo, and P. Vanska. “FASTAXON: A system for FAST (and Faceted) TAXONomy design.”. In *Procs. of 23th Int. Conf. on Conceptual Modeling (ER’2004)*, 2004. (an on-line demo is available at <http://fastaxon.erve.vtt.fi/>).
- [16] Y. Tzitzikas and C. Meghini. “Ostensive Automatic Schema Mapping for Taxonomy-based Peer-to-Peer Systems”. In *Procs. of the 7th Intern. Workshop on Cooperative Information Agents (CIA-2003)*, pages 78–92, 2003. (Best Paper Award).
- [17] Y. Tzitzikas, C. Meghini, and N. Spyrtatos. “A Unified Interaction Scheme for Information Sources”. *Journal of Intelligent Information Systems*, 26(1):75–93, January 2006.
- [18] Y. Tzitzikas, N. Spyrtatos, and P. Constantopoulos. “Mediators over Taxonomy-based Information Sources”. *VLDB Journal*, 14(1):112–136, 2005.
- [19] K. Yee, K. Swearingen, K. Li, and M. Hearst. “Faceted Metadata for Image Search and Browsing”. In *Proceedings of the Conf. on Human Factors in Computing Systems (CHI’03)*, pages 401–408, April 2003.