Revising Faceted Taxonomies and CTCA Expressions

Yannis Tzitzikas[1,2]

[1] Computer Science Department,
University of Crete, Greece
[2] Institute of Computer Science,
FORTH-ICS, Greece
tzitzik@csi.forth.gr

Abstract. A faceted taxonomy is a set of taxonomies each describing the application domain from a different (preferably orthogonal) point of view. CTCA is an algebra that allows specifying the set of meaningful compound terms (meaningful conjunctions of terms) over a faceted taxonomy in a flexible and efficient manner. However, taxonomy updates may turn a CTCA expression e ill-formed and may turn the compound terms specified by e to no longer reflect the domain knowledge originally expressed in e. This paper shows how we can revise e after a taxonomy update and reach an expression e' that is both well-formed and whose semantics (compound terms defined) is as close as possible to the semantics of the original expression e before the update.

1 Introduction

Let F be a faceted taxonomy, i.e. a set of taxonomies $(\mathcal{T}_1, \leq_1), \ldots, (\mathcal{T}_k, \leq_k)$, and let $\mathcal{T}=\mathcal{T}_1 \cup \ldots \cup \mathcal{T}_k$. Each expression e of CTCA (Compound Term Composition Algebra) [3] specifies a set S_e^F of valid (i.e. meaningful) compound terms (conjunctions of terms) over $\mathcal{T}$. So an expression e actually defines the partition $(S_e^F, \mathcal{P}(\mathcal{T}) - S_e^F)$ where $\mathcal{P}(\mathcal{T})$ denotes the powerset of $\mathcal{T}$. An update operation u_F on F (resulting to a faceted taxonomy F') may turn the expression e obsolete (i.e. not well-formed), or it may make the derived compound terminology $S_e^{F'}$ to no longer reflect the desire of the designer, i.e. it may no longer reflect the domain knowledge that was expressed in e. For example, the deletion of a term t may make several compound terms (that do not even contain t) to no longer belong to $S_e^{F'}$. It would be very useful if we could update automatically e to an expression e' that is (a) well-formed (w.r.t. F'), and (b) $S_{e'}^{F'}$ is as close to S_e^F as possible. This would enhance the robustness and usability of systems that are based on CTCA, like FASTAXON [2]. We call this problem *expression revision* after taxonomy update. In the ideal case, we would like to find an expression e' such as: (α) e' is well-formed, and $(\beta^{=})$ $S_{e'}^{F'} = S_e^F$. Although condition (α) can be satisfied quite easily, condition $(\beta^{=})$ may be impossible to satisfy in some cases. We can relax condition $(\beta^{=})$ and consider that our objective is to find an expression e'

G. Antoniou et al. (Eds.): SETN 2006, LNAI 3955, pp. 600–604, 2006.

such that $S_e^{F'}$ is as *close* to S_e^F as possible. We can define the distance between two compound terminologies S, S' as the cardinality of their symmetric difference, i.e.: $dist(S, S') = |(S - S') \cup (S' - S)| = |S - S'| + |S' - S|$. Now let $\mathcal{S}^{F'}$ be the set of *all* compound terminologies over F' that can be defined by expressions of CTCA. We can express condition (β) formally as follows: $S_{e'}^{F'} = \texttt{arg}_S\ \texttt{min}\{dist(S, S_e^F) \mid S \in \mathcal{S}^{F'}\}$. However, in some application scenarios, we may prefer $S_{e'}^{F'}$ to be a subset of S_e^F than being a superset, or the reverse. Consequently, we may state state two, different than (β), conditions: (γ) $S_{e'}^{F'} \subseteq S_e^F$ and $S_{e'}^{F'}$ is the biggest possible in $\mathcal{S}^{F'}$, and (δ) $S_{e'}^{F'} \supseteq S_e^F$ and $S_{e'}^{F'}$ is the smallest possible in $\mathcal{S}^{F'}$. Of course, to find the sought expression e' we would not like to investigate all expressions in $\mathcal{S}^{F'}$ (as this would be computationally inadmissible), but we rather want to find a method for *modifying* e to an e' that satisfies (α) and (β or γ or δ).

A complete treatment of this problem (with applications, examples, formal propositions and proofs) can be found at [1]. This paper is a short summary.

2 CTCA and Taxonomy Updates

The upper part of Table 1 recalls in brief the basic notions and notations about taxonomies and faceted taxonomies. Syntactically, a CTCA expression e over F is defined according to the following grammar ($i = 1, ..., k$): $e ::= \oplus_P(e, ..., e) \mid \ominus_N\ (e, ..., e) \mid \overset{*}{\oplus}_P\ T_i \mid \overset{*}{\ominus}_N\ T_i \mid T_i$. The initial operands, thus the building blocks of the algebra, are the *basic compound terminologies*, which are the facet terminologies with the only difference that each term is viewed as a singleton. In most practical settings, taxonomies have the form of trees and for reasons of space we confine ourselves to this case.

Plus-products and *minus-products*, denoted by $\oplus_P$ and $\ominus_N$ respectively, have a parameter that is denoted by P (resp. N) which is a set of compound terms over $\mathcal{T}$. In a P parameter the designer puts valid compound terms, while in a N parameter the designer puts invalid compound terms. The exact definition of each operation of CTCA (also including two auxiliary operations, called *product* and *self-product*) is summarized in the lower part of Table 1. An expression e is *well formed* iff every facet T_i appears at most once, and every parameter set P or N of e is always subset of the corresponding set of *genuine compound terms*. Specifically, the genuine compound terms in the context of an operation $\oplus_P(e_1, ..., e_k)$ (or $\ominus_N(e_1, ..., e_k)$) is denoted by $G_{e_1,...,e_k}$ and it is defined as: $G_{e_1,...,e_k} = S_{e_1} \oplus ... \oplus S_{e_k} - \cup_{i=1}^n S_{e_i}$.

We consider two update operations on subsumption relationships: $\texttt{delete}(t \leq t')$ (subsumption relationship deletion), and $\texttt{add}(t \leq t')$. Before $\texttt{delete}(t \leq t')$ we assume that the relationship $t \leq t'$ belongs to the transitive reduction (Hasse Diagram) of $\leq$, while before an operation $\texttt{add}(t \leq t')$ we assume that the relationship $t \leq t'$ does not already exist in $\leq$. We also consider three update operations on terms: $\texttt{rename}(t, t')$, $\texttt{delete}(t)$, and $\texttt{add}(t)$. Whenever a term t is deleted, all subsumption relationships in which t participates are deleted too,

Table 1. Notations and CTCA Operations

Name	Notation	Definition
terminology	$\mathcal{T}$	a finite set of names called terms
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	$F = \{F_1, ..., F_k\}$	$F_i = (\mathcal{T}_i, \leq_i)$, for $i = 1, ..., k$ 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 over S	$\preceq$	Given $s, s' \in S$, $s \preceq s'$ iff $\forall t' \in s' \ \exists t \in s$ such that $t \leq t'$.
immediate broaders of t	$Br_{(1)}(t)$	the smaller terms that are greater than t (w.r.t $\leq$), i.e. $minimal_{<}(\{t' \in \mathcal{T} \mid t \leq t', t \neq t'\})$
immediate narrowers of t	$Nr_{(1)}(t)$	the bigger terms that are smaller than t (w.r.t $\leq$), i.e. $maximal_{<}(\{t' \in \mathcal{T} \mid t' \leq t, t \neq t'\})$
broaders of t	$Br(t)$	$\{t' \in \mathcal{T} \mid t \leq t'\}$
narrowers of t	$Nr(t)$	$\{t' \in \mathcal{T} \mid t' \leq t\}$
broaders of s	$Br(s)$	$\{s' \in P(\mathcal{T}) \mid s \preceq s'\}$
narrowers of s	$Nr(s)$	$\{s' \in P(\mathcal{T}) \mid s' \preceq s\}$
broaders of S	$Br(S)$	$\cup\{Br(s) \mid s \in S\}$
narrowers of S	$Nr(S)$	$\cup\{Nr(s) \mid s \in S\}$
CTCA Operations		
Operation	e	S_e
	T_i	$\{ \{t\} \mid t \in \mathcal{T}_i\} \cup \{\emptyset\}$
product	$e_1 \oplus ... \oplus e_n$	$\{ s_1 \cup ... \cup s_n \mid s_i \in S_{e_i}\}$
plus-product	$\oplus_P(e_1, ...e_n)$	$S_{e_1} \cup ... \cup S_{e_n} \ \cup \ Br(P)$
minus-product	$\ominus_N(e_1, ...e_n)$	$S_{e_1} \oplus ... \oplus S_{e_n} - Nr(N)$
self-product	$\overset{*}{\oplus} \ (T_i)$	$P(\mathcal{T}_i)$
plus-self-product	$\overset{*}{\oplus}_P \ (T_i)$	$S_{T_i} \cup Br(P)$
minus-self-product	$\overset{*}{\ominus}_N \ (T_i)$	$\overset{*}{\oplus} \ (T_i) - Nr(N)$

however the other relationships are preserved, i.e. after deleting term t, for all $t' \in Nr_{(1)}(t)$ it holds $Br_{(1)}(t') \supseteq Br_{(1)}(t)$.

3 CTCA Expression Revision

Given a compound term s and a term t, we shall use $s\#t$ to denote the compound term $s - \{t\}$. Now given s and two terms t and t', we shall use $s\#t\#t'$ to denote the compound term s if $t \not\in s$, otherwise the compound term derived from s by replacing t by t'. We can generalize and define:

$$s\#s1\#s2 = \begin{cases} (s - s1) \cup s2, & \text{if } \ s \cap s1 \neq \emptyset \\ s & \text{otherwise} \end{cases}$$

We have studied expression revision for each kind of update operation and below we summarize the results. The deletion of terms or subsumption relationships can be handled by extending the P/N parameters (so as to recover the missing compound terms from the semantics of the original expression). Table 2 shows the revised expression e' after each taxonomy update. In cases (1),(2) and (4) Table 2 shows how each P and N parameter of e should be revised to a P' and N' parameter of e'. In case (3), for every minus-product operation $\ominus_N(e_1, ..., e_k)$ and for every e_i $(1 \leq i \leq k)$ such that $f(a) \not\in f(e_i)$

Table 2. Expression Revision after Taxonomy Updates

	u_F	alg	Notes
(1)	$\texttt{rename}(a, a')$	$P' = \{s\#a\#a' \mid s \in P\}$, $N' = \{s\#a\#a' \mid s \in N\}$ $S_{e'}^{F'} = \{s\#a\#a' \mid s \in S_e^F\}$	$\sim (\beta^=)$
(2)	$\texttt{delete}(a)$	$P' = \bigcup_{s \in P}\{s\#a\#t \mid t \in Br_{(1)}^F(a)\}$ $N' = \bigcup_{s \in N}\{s\#a\#t \mid t \in Nr_{(1)}^F(a)\}$ $S_{e'}^{F'} = S_e^F - \{s \mid a \in s\}$, thus $S_{e'}^{F'} \subseteq S_e^F$	(β)
(3)	$\texttt{add}(a)$	$N' = N \cup \{ \{a, u_i\} \mid e_i \in \text{operands}(cur_e),$ $f(a) \notin f(e_i), u_j \in maximal_{\preceq}(S_{e_i})\}$ $S_{e'}^{F'} = S_e^F \cup \{\{a\}\}$	$\sim (\beta^=)$
(4)	$\texttt{delete}(b \leq a)$	$P' = P \cup \{ s\#Nr^F(b)\#\{a\} \mid s \in P\}$ $N' = N \cup \{ s\#Br^F(a)\#\{b\} \mid s \in N\}$ $S_{e'}^{F'} = S_e^F$	$(\beta^=)$

(meaning the a belongs to a facet that does not appear in expression e_i), we have to add to N the parameter $\{a, u_i\}$ for each $u_i \in maximal_{\preceq}(S_{e_i})$. The most important result is that the addition of subsumption relationships 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 may neither be subset nor superset of the original compound terminology. This happens because the effects of adding a subsumption relationship is different in $\oplus_P$ and $\ominus_N$: the compound terminologies defined by $\oplus_P$ operations become larger, while those defined by $\ominus_N$ operations become smaller. Now the combination of $\oplus_P$ and $\ominus_N$ operations can lead to compound terminologies which are neither larger nor smaller than the original one. The following proposition gives sufficient and necessary conditions for satisfying condition $(\beta^=)$ after subsumption relationship addition.

Proposition 1. Let F' be the result of applying $\texttt{add}(b \leq a)$ on F. We can find an expression e' such that $S_{e'}^{F'} = S_e^F$ if and only if:

(i) for each $p \in P$ of every parameter P of e it holds:
If $p \cap Nr^F(b) \neq \emptyset$ then $\exists p' \in P$ such that $p' \preceq_F (p - Nr^F(b)) \cup \{a\}$
(ii) for each $n \in N$ of every parameter N of e it holds:
If $n \cap Br^F(a) \neq \emptyset$ then $\exists n' \in N$ such that $n' \succeq_F (n - Br^F(a)) \cup \{b\}$.

As a final remark we have to note that there is not any directly related work on the problem at hand because CTCA emerged relatively recently and its distinctive characteristics (range-restricted closed world assumptions) differentiate it from other logic-based languages (for more see [3]) and the corresponding literature on updates and revisions.

References

1. Y. Tzitzikas. “Updates and Revision in Faceted Taxonomies and CTCA Expressions”. Technical Report 2005-11-18, TR 364, Institute of Computer Science-FORTH, Nov 2005.
2. 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*, Shanghai, China, Nov. 2004.
3. Yannis Tzitzikas, Anastasia Analyti, and Nicolas Spyratos. “Compound Term Composition Algebra: The Semantics”. *LNCS Journal on Data Semantics*, 2:58–84, 2005.