Oponentský posudek na diplomovou práci

Martin Zeman
Využití ontologií pro GUHA procedury

Práce se zabývá využitím doménových znalostí ve formě ontologií v procesu dobývání znalostí z databází pomocí metody Guha. Cílem práce bylo implementovat moduly sloužící k zapojení ontologií do procesu zadávání úlohy v prostředí Ferda. V zadání práce se za součást řešení požaduje návrh ontologie, výběr vhodné reprezentace ontologie specifické pro potřeby GUHA procedur a experimentální ověření těchto modulů. Typ práce je podle klasifikace KSI stanoven na „implementační".

Hned na začátku je třeba říci, že se jedná o netriviální úlohu a jejíž rozsah patrně nebyl na začátku zřejmý (problém je totiž v tom že se jedná o implementaci, která je součástí širšího projektu, který byl vyvíjen v C# a ontologické nástroje jsou převážně v Java. I když .net technologie slibuje integraci různých platforem, konkrétně ve Ferdovi zvolené integrační prostředí ICE neposkytovalo přímočaré řešení a diplomant musel konzultovat osobně s tvůrci ICE). Konstatuji, že v implementaci, která je součástí širšího projektu – autor jasně vymezil zpracovanou část, včetně potřebného pojmového aparátu. V žádném případě se nejedná o rutinní implementaci uživatelské aplikace (viz požadavky KSI).

Navíc v sérii diplomek v prostředí Ferda je uchazeč jedním z prvních, kteří nebyly členy týmu, který Ferdu tvořil coby softwarový projekt (naštěstí vedoucí této DP byl členem tohoto týmu). Uchazeč se musel seznámit z uživatelským rozhraním systému Ferda (programátorská dokumentace) a celou teorií kolem GUHA-y (kapitola 1). V dalším musel nastudovat problematiku ontologií, jejich reprezentace a softwarové prostředky (kapitoly 2, 3 a 4). Takže se nejednalo o pouhou (i když netriviální) implementaci analyzovaného problému a navrženého řešení, ale i o nastudování minimálně dvou velikých teoretických oblastí.

Uchazeč provedl analýzu z různých hledisek implementace a navrhl řešení (i s diskusí alternativ). Při implementaci bylo patrně nejnáročnější integrovat do prostředí Ferda komponent parseru jazyka OWL, který byl napsán třetí stranou v jazyce Java. Užití API OWL parseru výrazně zlepšilo kvalitu řešení. Konkrétním výstupem je implementace modulu „Ontologie", modulu „Mapování ontologie" a modulu „Atribut odvozený z ontologie" jejich zaintegrování do „Ferdy" a propojení s externím úložištěm ontologií. Požadavek zadání „experimentální ověření těchto modulů" lze chápat dvojím způsobem – odladění SW nebo praktický přínos pro induktivní úlohu. První chápání lze považovat za splněné (až na drobné chyby), druhé asi potřebuje delší experimenty, na které patrně nezbyl čas.

Nebylo mi zcela jasné, proč něco (str. 43) nešlo modelovat s owl:Class, ale šlo to s rdfs:Class. Je problém v odlišnosti ontologií v Protegé, W3C specifikaci, a tou používanou v OWL API? Problém může být taky v hloubce modelování – kde končí třídy a kde začínají instance. Je to problém i pro mapování (nebylo by bývalo lepší zkvalitnit ontologii, která se mi zdá povrchní, nebo použít owl:subpropertyOf). Řešení, kde se tři vlastnosti kuřáků (délka, intenzita, ...) namapují na jednu vlastnost v ontologii (smoking) se mi nezdá nejšťastnější - možná by se mohl specifikovat vztah mezi nimi (rozdíl je stejný jako vztah mezi BMI a výškou a vahou, akorát že pro kouření ještě nikdo nenavrhl kombinaci těchto měr). Některé atributy ze Stulongu zůstávají pořád nesrozumitelné.

Nezmínil jste jiné alternativy OWL modelování rozdělení domén. Dalo se použít owl:oneOf ? Umí OWL API přistupovat k takovým datům.

Schází mi trochu kontrola toho, zda dělící body atributu jsou vůbec v jeho doméně (kde je typová kontrola, při tvorbě ontologie nebo až ve Ferda-ovi). Necitujete přístup z ČVUT – Železný, Žáková, který, sice v jiném kontextu ILP, používá informace o owl:subclassOf .

Myslím že i pro GUHA-u by to byla zajímavá informace, která by ušetřila prověřování spousty hypotéz.

Na závěr bych chtěl konstatovat, že předložená diplomová práce splňuje požadavky na implementační diplomovou práci a požadavky ze zadání (kromě „Rozsah tohoto zkoumání bude konzultován s vedoucím diplomové práce" – který jsem nedovedl posoudit).

Obsahuje několik novinek:

- jednak propojení ontologií a GUHA exploratační analýzy a
- přínos k nejpracnější části dolování z dat a to je příprava dat (80% času)
- možnost lepšího dělení domény na kategorie. Které je ve shodě s potřebou doménových specialistů (např. hraniční meze krevního tlaku).

Řešení přesahuje zadání, protože podporuje znovupoužitelnost, konkrétně, obsahuje návrh na vytvoření ontologií, v nichž je možné uchovávat rozšiřující informace o entitách ve formě hotové OWL šablony pro vytvoření takovýchto ontologií.

Práci M. Zemana doporučuji k obhajobě.

V Praze dne 24. 1. 2008

Prof. RNDr. P. Vojtáš, DrSc.