Oponentský posudok diplomovej práce Martina Ralbovského „Využití doménových znalostí při aplikacích GUHA procedur"

Hlavným cieľom predkladanej práce je navrhnúť spôsoby využitia vhodne reprezentovaných doménových znalostí pri dobývaní znalostí pomocou systému LISp-Miner. Doménové znalosti sú znalosti o oblasti, z ktorej sú analyzované dáta. V práci mali byť navrhnuté metódy na využitie takýchto znalostí pri definovaní atribútov a pri upresňovaní zadávania analytických úloh rôznych typov. Súčasťou práce mali byť počítačové experimenty využívajúce navrhnuté postupy.

Doménové znalosti uvažované v práci sú dvoch typov: ontológie a tzv. background knowledge. Ontológia je formálny popis konkrétnej vecnej oblasti typicky vo forme popisu pojmov a tried, vlastností jednotlivých pojmov a vzťahov medzi nimi. Background knowledge sú znalosti o danej vecnej oblasti v tvare asociačných pravidiel získaných od experta.

Autor uvažuje v práci oba typy doménových vlastností. Pre background knowledge autor zaviedol vlastnú formalizáciu vhodnú pre využitie v systéme Ferda, ktorý implementuje GUHA procedúry. Do systému Ferda zabudoval moduly na overovanie asociačných pravidiel background knowledge. Ako sám autor uvádza, tieto moduly sú vlastne filtrom hypotéz generovaných procedúrami 4FT a KL LISp-Mineru. Zobecnením je umožnenie abstraktných kvantifikátorov, čo sú v jeho prípade skupiny kvantifikátorov používaných v procedúre 4FT a dva nové kvantifikátory pre procedúru KL vyjadrujúce rastúcu a klesajúcu závislosť. Praktické použitie overovania pravidiel background knowledge urobil pre databázu STULONG.

Práca sa tiež teoreticky zaoberá využitím ontológií s ohľadom na ich možné využitie s GUHA procedúrami v rámci systému Ferda alebo v systéme Ever-Miner, ktorý zatiaľ existuje iba vo forme hrubého návrhu. Tu sa zaoberá možným využitím ontológií na identifikáciu chýbajúcich alebo naopak redundantných atribútov, tvorbu atribútov pomocou ontológií, tvorbu cedentov (vzorov podmienok pre GUHA procedúru) a tiež problému generovania úloh z ontológií. V tejto časti autor navrhol iba obecné postupy, ale neimplementoval ich. Celkovo sú uvedené úvahy v poriadku, ale často sa jedná o úvahy typu: Z ontológie zhromaždíme všetky atribúty (ako vlastnosti relevantných tried) a porovnáme tento zoznam so zoznamom stĺpcov tabuľky. S tým môžu byť problémy, ak niektoré stĺpce budú v ontológii reprezentované triedami, alebo inštanciami tried. Avšak ontológií existuje mnoho typov a úvaha končí doporučením, že by sa na to malo myslieť pri vytváraní ontológie.

Úvahy tohoto typu sú iste užitočné pre niekoho, kto bude tieto postupy zabudovávať do podobného systému, ale samotnej práci by asi prospelo, keby sa autor venoval viacej tomu čo implementoval. Myslím tým overovaniu background knowledge. Jeho experimentálne overovanie pravidiel background knowledge sa skladalo zo 16 úloh. Iba jedna dala kladný výsledok, že KL procedúra našla rastúcu závislosť medzi vzdelaním a mierou aktivity po zamestnaní. Ďalších 13 testov dopadlo negatívne. 3 úlohy skončili haváriou programu v časti, ktorú síce neimplementoval autor práce, ale je súčasťou systému Ferda, ktorého je pán Ralbovský spoluautorom. Pri vlastných experimentoch som zistil, že systém Ferda je značne nestabilný a pokusy uvádzané v diplomovej práci sa mi nepodarilo zreprodukovať.

Myslím, že tu je najväčšia slabina predkladanej práce. Keď autor zistil zaujímavé nové informácie – pravidlá, ktoré experti pre danú oblasť považujú za zrejmé, ale systém ich nepotvrdil – tak by bolo veľmi zaujímavé vedieť, s akou mierou spoľahlivosti tieto pravidlá platia, prípadne navrhnúť postupy ako stanoviť vhodný prah spoľahlivosti generovaných pravidiel. Vhodná miera spoľahlivosti by dovolila GUHA procedúram vygenerovať možno zaujímavé nové rovnako spoľahlivé pravidlá ako tie „obecne známe".

Text práce má drobné nedostatky v pravopise (napr. na poslednom riadku str. 47 chýbajú vo vete dve čiarky), ktoré však podstatne nekazia zmysel textu. Tiež som našiel pár nesprávnych odkazov (napr.

v poznámka 2 na st. 15 sa odkazuje na prácu [6], ale správne má byť [19]; časť 5.2.2 odkazovaná z predposledného riadku na str. 44 má byť 6.2.2).

Celkovo je práca pána Ralbovského prínosná a potvrdila charakter „výskumnej práce", keď sa v priebehu práce ukázalo, že bude užitočnejšie venovať sa background knowledge než viacej známym ontológiám. Implementácia navrhnutých postupov by si zaslúžila buď doladenie systému, do ktorého je vložená, alebo napojenie na LISp-Miner priamo. A aj autor sám naznačuje, že jeho práca by mohla byť základom ešte širšieho využitia doménových znalostí pri dobývaní znalostí s procedúrami GUHA. Preto doporučujem, aby práca pána Martina Ralbovského bola uznaná, ako diplomová práca.

Praha, 8.9.2006

RNDr. František Mráz, CSc.

KSVI MFF UK