Posudek oponenta diplomové práce

Jan Sochna:

Systém pro sběr XML dat a metadat z internetu

Cílem posuzované práce je vytvoření nástroje na prioritní získávání XML dokumentů z webu. Úloha je blízká standardní indexaci webu, je zde ale rozdíl v tom, že většina dokumentů na webu (html stránky) je zde využita pouze jako propojovací uzly sítě vedoucí k dokumentům ve formátech odvozených z XML (tedy i XSLT, XHTML, ...). Tyto pomocné dokumenty není třeba uchovávat. Další rozdíl je i v tom, že běžné crawlery pracují s dokumenty omezené délky (běžné html stránky mívají jen omezenou velikost), zatímco některé XML dokumenty mohou obsahovat rozsáhlé soubory dat. Daného cíle tak není možné dosáhnout přímo využitím stávajících webových crawlerů.

Text práce je stručný, po jazykové i typografické stránce na odpovídající úrovni. Přiložené CD obsahuje zdrojové kódy i live verze obou částí (stahovače i analyzátoru) i startovní data pro stahování. Návod na zprovoznění je příliš stručný, aplikaci se mi nepodařilo zprovoznit. Vyžádal jsem si proto asistenci autora. Po jistém úsilí se autorovi podařilo dílo v plném rozsahu předvést. Z dalších připomínek:

- Ani v práci ani na CD není určeno, s jakými verzemi operačního systému a pomocného softwaru dílo spolupracuje.
- Návod na zprovoznění předpokládá netriviální znalosti a úsilí – přesnější návod by pomohl snazšímu přijetí díla.
- Není přesně řečeno, co který parametr vstupu znamená, ani co přesně uživatel může dostat na výstupu.
- Některá zajímavá rozhodnutí nejsou zdokumentována. Ti, kdo by chtěli na práci navázat, tak musí celou věc přezkoumat sami znovu.
- Je obtížně určitelné, co všechno je dílem diplomantovým a co bylo převzato.
- Seznam literatury není uspořádán ani dle abecedy, ani dle prvního výskytu.
- Projektu by prospěla možnost přesně (pozitivně i negativně) omezit oblast, kde se smí vyhledávat.

Odevzdaná práce i přes výše uvedené připomínky splňuje nároky kladené na diplomovou práci, doporučuji ji proto k obhajobě.

Praha, 19. května 2010

RNDr. Michal Žemlička, Ph.D.