INSTITUTE OF COMPUTER SCIENCE
Academy of Sciences of the Czech Republic
Martin Holeňa

Pod Vodárenskou věží 2, 182 07 Praha 8, phone: +420 266052921, fax: +420 286585789, e-mail: martin@cs.cas.cz
web: www.cs.cas.cz/~martin

Posudek habilitační práce

Interpretable Data Analysis with Entity-based Text Representations and Rule-based Models

Tomáše Kliegra

Habilitační prací Tomáše Kliegra je komentovaný soubor sedmi publikací. Chtěl bych vyzdvihnout, že z těchto 7 publikací je 5 časopiseckých. Přitom zvláště vysoce vyzdvihnout bych chtěl to, že tři z nich vyšly v časopisech, které v době vydání byly z hlediska impaktního faktoru ve druhém decilu, přičemž časopis Knowledge-Based Systems, v němž vyšly dva články, byl vždy na samém začátku druhého decilu. Ale i zbývající dvě časopisecké publikace měly impaktní faktor kolem mediánu, takže vzhledem k nárokům kladeným na časopisecké články se i v jejich případě rozhodně jedná o publikace nadprůměrné kvality. A totéž lze říci i o dvou konferenčních publikacích ze souboru. Obě byly na konferencích zahrnutých do seznamu informatických konferencí CORE (Computer Research and Education), přičemž jedna z nich je v něm vedena jako konference kategorie A a druhá jako konference kategorie B.

První čtyři publikace ze souboru se týkají strojově zpracovatelné reprezentace textu pomocí entit. Jako klíčovou vnímám publikaci „Linked hypernyms: Enriching DBpedia with Targeted Hypernym Discovery“ vyšlou v časopise „Web Semantics: Science, Services and Agents on the World Wide Web”, která prezentuje rozšíření sémantických znalostních bází DBpedia a YAGO nazvané Linked Hypernyms Dataset (LHD). Velký dojem na mě udělalo, že pod vedením Tomáše Kliegra byl LHD naplněn téměř 5 milióny přiřazení typů entitám na základě hesel z anglické, německé a nizozemské mutace Wikipedie. Tento přínos LHD byl oceněn DBpedia TextExt Challenge Prize. Zdaleka ale nejde jen o kvantitu, experimentální výsledky prezentované ve výše uvedené publikaci ukazují i velkou přesnost přiřazení obsažených v LHD, a to ve všech třech jazycích.

Z ostatních výsledků v první části práce si nejvíce cením toho, že pravidla naučená na interpretovatelných entitách v konečném důsledku umožňují interpretovatelnou klasifikaci dokumentů.

Poslední tři publikace se týkají získávání asociačních pravidel z dat. Z nich vnímám jako klíčový článek „Web framework for interpretable machine learning based on rules and

frequent itemsets“ v časopise „Knowledge-Based Systems“, který je nejdůležitější publikací o systému EasyMiner. Nejvíce si cením toho, že snaha o srozumitelnost znalostí získaných z dat se v systému EasyMiner neomezila jen na použití asosicačních pravidel jakožto reprezentace znalostí srozumitelné člověku, ale zahrnula i grafické rozhraní systému a také klestění množin pravidel získaných dat, které je obranou proti jejich přílišné početnosti, znehodnocující srozumitelnost jednotlivých pravidel. Velký dojem na mě také udělala souvislost systému EasyMiner s detekcí anomálii, vzhledem k tomu, že v teorii detekce anomálií není věnována téměř žádná pozornost reprezentaci anomálií pomocí člověku srozumitelných pravidel, přestože ve dvou hlavních aplikačních oblastech detekce anomálií, detekci malware a detekci útoků v síti, hraje srozumitelnost získaných znalostí o anomáliích pro člověka, konkrétně jejich srozumitelnost pro analytiky malware a pro operátory sítí, mimořádně důležitou roli. Kromě toho je třeba vyzdvihnout flexibilitu systému EasyMiner, který umožňuje začlenění různých algoritmů pro získávání asociačních pravidel z dat, jak nejrozšířenějších algoritmů Apriori a Frequent-Pattern-Growth, tak i algoritmů implementovaných v systému LISP-Miner, určených pro obecnější třídu pravidel.

Z ostatních výsledků ve druhé částí práce si nejvíce cením použití asociačních pravidel jako vysvětlující komponenty doporučovacího systému, a to navíc sytému, který se úspěšně zúčastnil dvou doporučovacích soutěží v letech 2013 a 2014, v nichž se musel vypořádat s velkými objemy dat a s požadavkem na krátkou dobu odezvy.

Komentář k soubor publikací tvořících habilitační práci zasvěceně zasazuje tyto publikace do vývoje obou oblastí, jichž se týkají, jakož i do celkové situace výzkumu a výuky na VŠE. Postrádám v něm pouze jednu drobnost: při diskuzi o rozdílu mezi metodami získávajícími z dat znalosti reprezentované pomocí pravidel a umělými neuronovými sítěmi by si zasloužily zmínku alespoň některé z řady specializovaných metod vyvinutých od konce 80. let pro extrakci pravidel z natrénovaných neuronových sítí. Jsou zmíněny pouze dvě metody pro extrakci pravidel z jakýchkoliv klasifikátorů, nicméně specifické metody pro neuronové sítě lze použí nejen při používání sítí ke klasifikaci, ale při jejich – díky univerzální aproximační schopnosti umělých neuronových sítí pravděpodobně mnohem známějším – používání k regresi.

Na závěr bych rád zdůraznil, že předloženou habilitační práci jednoznačně doporučuji k obhajobě. Publikace, které ji tvoří, ukazují Tomáše Kliegra v obou oblastech, kterých se týkají, jako vědeckou osobnost mezinárodní úrovně. Zejména ve druhé z nich, získávání asociačních pravidel z dat, pokládám Tomáše Kliegra za důstojného pokračovatele světově věhlasné české školy získávání srozumitelných pravidel z dat, u jejíhož zrodu stál ve 2. polovině 60. let Petr Hájek, a k jejímuž rozvoji přispěla od té doby řada dalších českých matematiků a informatiků, asi nejvíce Tomáš Havránek, Vilém Novák a starší kolega Tomáše Kliegra na VŠE Jan Rauch.

9. ledna 2019

Martin Holeňa