

V ý p i s

ze zápisu ze zasedání vědecké rady Fakulty informatiky a statistiky VŠE v Praze,
konané dne 14. 2. 2019

Přítomni: dle prezenční listiny

Program: ad 4) Habilitační řízení **Ing. Tomáše Kliegra, Ph.D.**, docentem pro obor Aplikovaná informatika

Děkan Fakulty informatiky a statistiky prof. Ing. Jakub Fischer, Ph.D. seznámil členy vědecké rady se složením habilitační komise, která posuzovala materiály k habilitačnímu řízení Ing. Tomáše Kliegra, Ph.D. Předsedou habilitační komise byl jmenován prof. Ing. Josef Basl, CSc., členy komise byli prof. Ing. Petr Berka, CSc., prof. RNDr. Jaroslav Pokorný, CSc., doc. RNDr. Lubomír Popelínský, Ph.D. a prof. Ing. Filip Železný, Ph.D.

Jako skrutátoři byli přítomnými členy vědecké rady schválení navržení kandidáti, a to doc. Ing. Klára Antlová, Ph.D. a prof. RNDr. Josef Hynek, MBA, Ph.D.

Poté se ujal slova předseda komise prof. Basl, který v úvodu svého vystoupení představil přítomným kandidáta a uvedl jeho základní odbornou charakteristiku, která je podrobně rozvedena ve Zprávě habilitační komise. Zdůraznil, že komise důkladně zhodnotila pedagogickou, vědecko-výzkumnou a publikační činnost kandidáta. Komise konstatuje, že všechny tři oponentské posudky habilitační práce jsou jednoznačně kladné a všichni oponenti práci doporučují k obhajobě. Komise konstatuje, že Ing. Tomáš Kliegr, Ph.D. splňuje všechny požadavky kladené na uchazeče o titul *docent* uvedené v Kritériích uplatňovaných při habilitačním řízení a řízení ke jmenování profesorem na Vysoké škole ekonomické v Praze.

S ohledem na jeho odborné zaměření Komise navrhl, aby žadatel přednesl habilitační přednášku na téma: *Klasifikace na bázi asociačních pravidel*.

Po úvodním slovu předsedy komise vyzval děkan fakulty Ing. Kliegra k přednesení habilitační přednášky na uvedené téma.

V úvodu habilitační přednášky nejprve kandidát konkretizoval cíl přednášky, dále stručně uvedl úlohu dobývání asociačních pravidel v historickém kontextu. Kandidát následně představil algoritmus Apriori jako nejznámější přístup pro dolování asociačních pravidel. Úspěchy tohoto algoritmu poté motivoval jeho adaptaci pro klasifikační úlohu. Hlavní část přednášky byla věnována popisu algoritmu Classification based on Associations (CBA). Po stručném představení dat použitých k ilustraci výkladu kandidát postupně popsal hlavní kroky algoritmu: (1) diskretizaci, (2) dolování častých množin položek, (3) dolování asociačních pravidel a (4) tvorbu klasifikátoru z nalezených pravidel a následné použití modelu pro predikci.

Po dokončení výkladu vztahujícímu se k algoritmu CBA, přešel kandidát k druhému tématu přednášky – charakteristice dalších příbuzných algoritmů a stručnému srovnání různých přístupů pro klasifikaci z pohledu rovnováhy mezi srozumitelností modelů a správností klasifikace. V závěru přednášky kandidát popsal charakteristické rysy klasifikace pomocí asociačních pravidel a uvedl praktické obtíže při použití algoritmu CBA.

V diskusi byly předneseny tyto dotazy:

prof. Pokorný: v habilitační přednášce byla vidět ta nejjednodušší pravidla v antecedentu například s konjunkcí. Jak je to například s negací? Jsou někde postupy, že v konjunkci může být negace?

Odpověď kandidáta. Problémem využití expresivnějších asociačních pravidel na vstupu algoritmu CBA jsme se zabývali. Provedli jsme řadu benchmarků s použitím pravidel generovaných metodou GUHA, které umožňují negace i disjunkce mezi hodnotami jednoho atributu. Ukázalo se, že expresivnější jazyk příliš nepomohl. Výsledky této předběžné studie jsou zahrnuté v habilitační práci.

prof. Pokorný: jak se hledání asociačních pravidel chová v případě Big Data?

Odpověď kandidáta. První fáze algoritmu CBA je na to připravena díky použití algoritmu Apriori, který byl navržen právě pro zpracování velkých dat. Navíc algoritmus Apriori lze v rámci CBA nahradit i dalšími algoritmy (FP-Growth), které mohou přinést další zrychlení. Evaluaci celého algoritmu CBA jsme se věnovali v rámci grantu od Fondu rozvoje CESNET, kdy se nám podařilo zpracovat data přibližně o velikosti 1 GB.

prof. Černý: jaká je složitost algoritmu? Vypadá to, že se jedná o exponenciální složitost.

Odpověď kandidáta. Přesnou odpověď na to nemám. Velmi záleží na konkrétních datech a nastavení minimálního prahu podpory. Algoritmus dokáže efektivně redukovat prostor prohledávání za využití tzv. vlastnosti Apriori, která spočívá v následujícím: všechny podmnožiny množiny častých položek splňujících uživatelem zadanou minimální podporu musí také splňovat minimální podporu. V praxi tedy např. pro kandidátní množinu položek o délce 2 musí všechny její podmnožiny obsahující jednu položku také splňovat minimální podporu. Tímto způsobem dochází obvykle k rychlé redukci počtu kombinací. Může se ale stát, že pro konkrétní data a určitou hranici podpory vlastnost Apriori rychlou redukci počtu kombinací nezajistí, a dojde ke „kombinatorické explozi“ náročné jak výpočetní výkon, tak na paměť.

Ing. Svoboda: klasifikátor na bázi asociačních pravidel je velmi zavedený. Vedle toho jsou neuronové sítě se svými úspěchy, ale současně jsou v počátku. Jak vidíte budoucnost neuronových sítí a CBA z hlediska správnosti?

Odpověď kandidáta. V současnosti je společenská poptávka po interpretovatelných modelech. V rámci soutěže Explainable Machine Learning Challenge organizované společností FICO bylo vzorové řešení poskytnuto formou vysvětlení připomínajících pravidla. Uvedené ilustruje to, že například regulátoři mohou preferovat přímo interpretovatelné modely, jako jsou pravidla. „Black-box“ modely, jako jsou neuronové sítě, dosahují ale lepších výsledků.

Ve druhé části vědecké rady obhajoval Ing. Kliegr předloženou habilitační práci, kterou tvoří souhrn komentovaných publikací s názvem *Interpretable Data Analysis with Entity-based Text Representations and Rule-based Models*. Úvodem přiblížil kandidát téma habilitační práce, její cíle a rámcové členění. Následně poukázal na souvislosti s habilitační přednáškou v oblasti motivace, kdy pravidlové modely dosahují velmi dobré interpretovatelnosti.

Poté kandidát podrobněji představil náplň každého ze sedmi příspěvků zahrnutých do habilitační práce. U každého příspěvku zdůraznil nejdůležitější výsledky a místy poukázal na možnosti využití či aplikací v praxi.

Po ukončení obhajoby zhodnotil habilitační práci i přítomný oponent prof. Holeňa.

prof. Holeňa: oponent se stručně vyjádřil ke kvalitě publikací zahrnutých v souboru. Několik ze zahrnutých časopiseckých příspěvků je vedených v Q1 a zbylé jsou také v přiměřeně kvalitních časopisech. Dva zahrnuté konferenční příspěvky byly publikované na konferencích dosahujících dobrého umístění v rejstříku CORE. Prof. Holeňa dále zasadil práci kandidáta do kontextu pražské školy věnující se již od 60. let 20. století objevování pravidel v datech, přičemž jmenoval již zesnulého prof. Hájka a doc. Havránka, a dále prof. Raucha z KIZI VŠE. Praktický dopad práce pak byl uveden kandidátem během obhajoby práce. Závěrem vyslovil doporučující stanovisko pro obhajobu habilitační práce a jmenování kandidáta docentem.

S posudky omluveného doc. Smrže a prof. Vojtáše seznámil členy vědecké rady předseda habilitační komise prof. Basl.

doc. Smrž - prof. Basl stručně shrnul text posudku a přečetl doporučující závěr: „I can ascertain that the reviewed habilitation thesis proves author's excellent research track in the knowledge engineering field and clearly demonstrates his potential for future scientific work in the area. I propose to accept it for the habilitation at the University of Economics, Prague.“ Následně prof. Basl shrnul připomínky v posudku, kdy upozornil na vynechání některých aktuálních metod z úvodní rešerše. Konkrétně uvedl metody používané pro vysvětlení neuronových sítí, problematiku kontextové podobnosti slov (metoda BERT) a systém DeepType.

prof. Vojtáš - prof. Basl stručně shrnul text posudku a přečetl doporučující závěr posudku: „The results of this thesis are important, published in good journals and deal with hot topics and hard problems... I approve the work as a habilitation thesis.“ Následně prof. Basl přednesl otázku obsaženou v posudku směřující na uživatelské studie důvěryhodnosti pravidel.

Všechny tři posudky byly kladné a doporučující.

V rozpravě k habilitační práci nejdříve kandidát reagoval na připomínky z posudků oponentů. V reakci na připomínku vztahující se k extrakci pravidel z neuronových sítí (prof. Holeňa) a dalších metod pro interpretaci kvantitativních modelů (doc. Smrž) kandidát uvedl, že práce se zaměřovala především na přímo interpretovatelné modely, což ale z úvodu práce nebylo zřejmě jasné. Zaměření na přímo interpretovatelné metody mělo být více akcentováno. S metodami pro extrakci pravidel z neuronových sítí se kandidát seznámil v rámci stáže na TU Darmstadt, kde rozvíjí metodu DeepRED. Mezi jiné algoritmy pro vysvětlení neuronových sítí patří LIME a surrogate decision trees.

Dále kandidát reagoval na připomínky doc. Smrže vztahující se k vynechání některých aktuálních metod vztahujících se k práci s entitami. Kandidát uvedl, že metoda DeepType byla publikována v roce 2018 krátce před odevzdáním práce, systém BERT až po jejím odevzdání. Existovaly ale předchozí příspěvky, na kterých jsou tyto metody založeny. Systém BERT vychází z metody Deep contextualized word representations (Peters, 2018), která byla publikována ještě před odevzdáním práce. Kandidát si ji nepovšiml, nejspíš kvůli nižšímu počtu citací, které měla v době dokončování habilitační práce. Kandidát se v současnosti soustřeďuje více na problematiku učení pravidel, poslední vývoj v oblasti rozpoznání entit mu mohl uniknout. K dotazu prof. Vojtáše kandidát uvedl, že na toto téma mu se zahraničním kolegou krátce po odevzdání práce vyšel příspěvek „The need for interpretability biases“ (Springer, 2018) a další dva jsou k dispozici na preprint serveru arXiv. K připomínce vztahující se k problému snižování komplexity pravidel kandidát uvedl, že se problémem aktuálně zabývá a odkázal na příspěvek s pracovním názvem Quantitative CBA, který je k dispozici na preprint serveru arXiv.

Následně otevřel děkan fakulty veřejnou rozpravu, v níž vyzval členy vědecké rady ke vznesení dotazů a připomínek.

prof. Pokorný: pravidel je většinou velké množství (desetitisíce) a nemáme prostor k tomu se ujistit, že pravidla jsou konzistentní. V prostředí webu je mnoho nekonzistentních dat, což výrazně stěžuje komplexní přístup k analýze srozumitelnosti.

Odpověď kandidáta. V aktuální práci se zaměřujeme na srozumitelnost jednotlivých pravidel. Hodnocením srozumitelnosti celého klasifikátoru jsme se zatím nezabývali právě z důvodu, že to považujeme za velmi obtížný problém.

prof. Cipra: v současné době se řeší etika strojového učení. Například, zda některé texty mohou mazat roboti. Václav Klaus mladší navrhuje nic nemazat, naopak komisařka Jourová chce ponechat možnost mazání, ale mohli by to dělat jen lidé. Jak to vidíte?

Odpověď kandidáta. Podobné téma jsme diskutovali v rámci přípravy jednoho projektu s odborníky z několika veřejnoprávních médií, včetně jednoho zahraničního. Zastávali názor, že by tam měl být nějaký kvalifikovaný moderátor odlišný od státu i koncových uživatelů. Strojové učení nalezne kandidáty na nevhodný obsah a odborníci to pak posoudí. Výsledkem posouzení není vymazání příspěvku, ale uvedení relevantních informací, které – například – danou dezinformaci vyvracejí.

prof. Cipra: dat ale je velké množství.

Odpověď kandidáta. Velká data lze řešit pomocí shlukování textů a algoritmů pro hluboké porozumění textu (jako Deep QA). Odborníci se následně věnují jen celým shlukům. Existují tedy technologické alternativy k cenzuře.

prof. Fiala: oponenti měli připomínky, že v habilitační práci něco chybí. Jaký je ale Váš největší vlastní přínos?

Odpověď kandidáta. Kandidát popsal jednotlivé příspěvky zahrnuté v práci. S ohledem na časové omezení shrnul svůj přínos u vybraných částí habilitační práce následně: provedení evaluací a benchmarků parametrů algoritmu CBA, pokus o návrh nové metody kombinující expresivnější pravidla GUHA s algoritmem pro tvorbu klasifikátoru v rámci metody CBA, návrh nových algoritmů pro extrakci typů z textu a návrh nové metodiky a datasetů pro evaluaci metod pro výpočet sémantické podobnosti slov. Odkázal také na vymezení příspěvků spoluautorů.

prof. Brezina: jaké vlastnosti musí znalostní graf v kontextu teorie grafů splňovat, abychom jej tak mohli nazvat?

Odpověď kandidáta. Znalostní graf je pojem, který je obvykle používán mimo kontext teorie grafů. Znalostní graf je tvořen uzly, které jsou spojeny orientovanými hranami. Na graf nejsou typicky kladeny další požadavky ve smyslu teorie grafů.

prof. Pokorný: v čem je to tedy znalostní?

Odpověď kandidáta. Grafy jsou obvykle založeny na standard W3C RDF, kdy pro výchozí uzel se používá pojem subjekt, pro cílový uzel pojem objekt, a pro hranu pojem predikát. Uzly hrany jsou pojmenované.

Na všechny dotazy a připomínky odpověděl Ing. Kliegr ke spokojenosti všech tazatelů a členů vědecké rady.

Po ukončení této části proběhla neveřejná diskuse.

Po ukončení této části vyzval děkan fakulty v neveřejné části zasedání členy vědecké rady k tajnému hlasování, jehož výsledek je následující:

- počet členů vědecké rady: **42**
- počet členů VR přítomných: **30**
- počet odevzdaných hlasů: **30** kladných
0 neplatných
0 záporných.

Děkan Fakulty informatiky a statistiky Vysoké školy ekonomické v Praze předloží podle § 72 odst. 11, zákona č. 111/1998 Sb. rektorce Vysoké školy ekonomické v Praze návrh na jmenování

Ing. Tomáše Kliegra, Ph.D.
d o c e n t e m
pro obor Aplikovaná informatika

Zapsala: Ing. Marie Gvoždiaková

Schválil: prof. Ing. Jakub Fischer, Ph.D.
děkan FIS