# Oponentský posudok dizertačnej práce

Doktorand: **Ing. Martin Dostal**

Téma dizertačnej práce: **Text-mining with linked data**

Pracovisko: **Fakulta aplikovaných věd, Západočeská univerzita v Plzni**

Predložená dizertačná práca Ing. Martina Dostala sa zaoberá v prvom rade viacerými spôsobmi, ako využiť sémantickú informáciu z prepojených dát (angl. Linked Data) pre riešenie úloh využívajúcich dolovanie v textoch. V druhom rade je časť práce venovaná aj využitiu algoritmu PageRank a jeho modifikácií pre extrakciu príznakov v kontexte klasifikácie textov a citačnej analýzy. Vo všetkých prípadoch ide o aktuálnu a vedecky zaujímavú problematiku spadajúcu do odboru Informatika a výpočetní technika, v ktorom sa autor uchádza o titul PhD.

Po preštudovaní dizertačnej práce konštatujem, že doktorand splnil všetky stanovené ciele dizertačej práce (bolo ich spolu 5). Aj keď cieľ G2 (aplikácia a vyhodnotenie metód navrhnutých v rámci cieľa G1) považujem za metodicky nevyhnutný a považujem ho skôr za súčasť cieľa G1. K jeho vyčleneniu od cieľa G1 autora zrejme viedla skutočnosť, že navrhnuté metódy extrakcie príznakov použil v dvoch rôznych úlohách, a to zhlukovanie, resp. klasifikácia textov. Zvolené metódy spracovania, ktoré pritom použil, sú adekvátne charakteru stanovených cieľov. Všetky navrhnuté metódy boli vhodným spôsobom experimentálne overené, aj keď v niektorých prípadoch by si zaslúžili výsledky experimentov podrobnejší rozbor a najmä väčšiu snahu porovnať ich s adekvátnymi prístupmi iných autorov.

- Za veľmi zaujímavý výsledok považujem aj postup pre extrakciu kľúčových fráz, ktorý autor popisuje v podkapitole 3.3 (mimochodom na publikáciu popisujúcu tento postup z Dateso 2011 sú už viaceré ohlasy). Vykonané experimenty vykazujú veľmi zaujímavé výsledky. Nechýba ani porovnanie s inými autormi, aj keď vykonané experimenty neboli na rovnakých dátových množinách. Nie je mi celkom jasné, prečo tento zaujímavý výsledok nebol zahrnutý medzi ciele dizertačnej práce.
- Za najlepšie spracovanú považujem kapitolu 4, ktorá prináša aj dva dôležité výsledky (spĺňajúce ciele G1, G3 a čiastočne aj G2). Autorom navrhnutá metóda extrakcie príznakov založená na prepojených dátach dáva veľmi dobré výsledky. Škoda že autor neporovnáva svoje výsledky s inými autormi.
- Veľmi zaujímavá je aj ďalšia metóda navrhnutá autorom na popis zhlukov (časť 4.3, cieľ G3) využívajúca grafovú štruktúru prepojených dát. Metóda by si určite zaslúžila viac experimentov než iba ten prezentovaný v práci a taktiež porovnanie s inými autormi. Tie isté pripomienky sa týkajú aj autorom navrhnutej metódy pre extrakciu príznakov za pomoci prepojených dát, s využitím algoritmu PageRank, vyhodnocovanom na úlohe klasifikácie textov (kapitola 5, čiastočne ciele G1 a G2).
- Za originálny považujem nápad využiť prepojené dáta na analýzu textov softvérových špecifikácií (kapitola 6, cieľ G5). Oceňujem že autor si za týmto účelom vytvoril aj vlastný korpus, na ktorom dosiahol veľmi sľubné výsledky experimentov.

- Autor sa podieľal aj na prestížnom výsledku publikovanom vo vysoko impaktovanom časopise Journal of Informetrics, ktorý je popísaný v kapitole 7 (cieľ G5). Svoj podiel na tomto výsledku vysvetľuje hneď v úvode, v časti 1.4.

Z formálneho hľadiska je predložená dizertačná práca na dobrej úrovni, autor sa myslím celkom úspešne popasoval aj s angličtinou. Nie celkom ideálne pôsobí však jej štruktúra, kde autor v snahe uviesť ciele dizertačnej práce najprv uvádza stručný úvod do problematiky v prvej kapitole. Viaceré informácie z tejto kapitoly sa potom opakujú neskôr v kapitolách popisujúcich naplnenie jednotlivých cieľov. K práci mám zopár menších pripomienok a otázok:

- U vzťahu (1.2) na str. 4 nie je uvedený význam veličiny $S_{ij}$. Tiež v nasledujúcom vzťahu (1.3) nie sú vysvetlené použité veličiny.
- Autor v kapitole 5 na str. 53 uvádza ako relevantné práce pomerne staršieho dáta. Nie sú známe aj novšie prístupy?
- V zhodnotení výsledkov rôznych spôsobov výpočtu PageRank skóre (Tabuľka 5.1) autor uvádza jednoznačné tvrdenie, že najlepší je variant c). Aké kritérium pritom použil?
- Na str. 75 je uvedené tvrdenie, že vysoký počet publikácií môže identifikovať populárnych autorov. Je to naozaj v počte publikácií, nie skôr v počte ohlasov na ne?

Publikačnú činnosť doktoranda považujem za veľmi kvalitnú, nakoľko doktorand už má štyri časopisecké publikácie (z toho jeden karentovaný), pričom v dvoch z nich je prvým autorom, šesť konferenčných príspevkov (z toho 4 sú indexované v renomovaných databázach). Doktorandova vedecká práca priniesla už aj päť ohlasov.

Na základe uvedeného hodnotenia predloženú dizertačnú prácu Ing. Martina Dostala odporúčam k obhajobe a súhlasím aby mu bola udelená vedecko-akademickú hodnosť Philosophiae doctor (PhD.).

V Košiciach, 21. januára 2015

prof. Ing. Ján Paralič, PhD.
Katedra kybernetiky a umelej inteligencie
Technická univerzita v Košiciach
Letná 9, 042 00 Košice
Slovenská republika

# Posudek dizertační práce

**Dizertační práce**
Text-mining with linked data
*Text-mining s využitím linked data*
**Instituce**
Západočeská univerzita v Plzni, Fakulta aplikovaných věd
**Autor**
Ing. Martin Dostal
**školitel**
prof. Ing. Karel Ježek, CSc.
**Oponent**
doc. RNDr. Petr Šaloun, Ph.D.

## Přehled

Úvod dizertační práce obsahuje motivaci a stav poznání zejména v základních oblastech klastrování a klasifikace, a dále uvádí, mimo jiné, cíle práce. Druhá kapitola popisuje přehled současného stavu poznání v oblastech sémantického webu a propojených dat. Kapitola třetí uvádí metody extrakce klíčových slov, které autor rozšiřuje o vlastní přístupy k získávání klíčových slov a slovních spojení. Zde je uveden první návrh využití propojených dat jako zdroje sémantické informace. Kapitola čtvrtá je věnována clusterování od základní definice k autorovým vlastním metodám vytěžování dat z textu, zejména zjišťování označení klastrů s využitím informací získaných prostřednictvím propojených dat. Kapitola pátá se popisuje algoritmus klasifikace od definice k experimentům s výběrem vlastností, při kterém spojuje PageRank a propojená data s cílem určení nejvýznamnějších vlastností dokumentu. Kapitola šestá zavádí nový přístup k sémantické analýze specifikací softwaru. Svým přístupem doktorand kombinuje a spojuje různé veřejné a soukromé zdroje dat. Kapitola sedmá shrnuje experimenty realizované s dalšími členy výzkumné skupiny zaměřené na PageRank a citační sítě. Cíle práce jsou uvedeny v úvodu práce a následně přehledně shrnuty v jejím závěru, a to včetně odkazů na části dizertace, kde jsou diskutovány.

Dizertační práce je vypracována v oblasti webového inženýrství, kam lze vytěžování dat s využitím propojených dat zařadit. Spojení uvedených oblastí výzkumu je velmi aktuální, při ověření výsledků mohly být využity dostupné datové sady, navíc doktorand připravil i několik vlastních datových sad, které nabízí k využití odborné veřejnosti. V práci jsou navrženy úpravy známého PageRank a jsou diskutovány experimenty v oblasti scientometriky. Práce je napsána anglicky, text je dobře srozumitelný a jednotlivé části dizertace na sebe správně a logicky navazují.

## Aktuálnost zvoleného tématu

Dizertační práce představuje výsledky původního aktuálního výzkumu v oblasti, která se stále rozvíjí. Možnost praktického využití výsledků výzkumu doktoranda naznačuje podpora scientometrického jeho výzkumu významným mezinárodním časopisem.

## Poznámky k textu a diskuze

Rád konstatuji, že s původními výsledky doktoranda i jejich prezentací jsem se osobně seznámil již v roce 2012 na konferenci Datakon.

U obhajoby uvítám diskusi na témata:

- Jaké spatřujete možnosti využití Vašich výsledků při detekci plagiátů?
- Analyzoval jste datovou sadu CFP různých konferencí, dokázal byste navrhnout pomocí Vašich výsledků předzpracování příspěvků s ohledem na CFP?

Některé zkratky jsou použity a nejsou definovány, např. TF-IDF, IR. Nekonzistentně jsou formátovány položky seznamů, kdy jsou položky seznamu ukončovány čárkou (strana 8), nebo nejsou ukončeny vůbec (strana 17), současně se na jedné straně 4 vyskytuje obojí. V anglicky psaném textu bych přitom upřednostnil ';'. Na straně 13 se doktorand správně odkazuje na koncept Linked data, totéž v souvislosti se Semantic web postrádám. Oceňuji značení mnoha URL jako po známky pod čarou, neboť jejich charakter takovému umístění odpovídá. Zarovnání číselných hodnot na střed ve sloupcích tabulek, považuji za nevhodné, podobně, jako jejich zvýraznění podtržením – vše např. v tabulce 4.2 na str. 40. Průměrování hodnot přes různé testy je bez dalšího rozboru nevhodné, myslím např. tabulky 4.1 až 4.3, rozdělení tabulky zalomením stránky není u malých tabulek rovněž vhodné.

## Shrnutí

Práce obsahuje původní výsledky výzkumu a odkazuje se i na další výsledky doktoranda, uvedené v předběžné verzi tezí (thesis proposal). Výsledky doktoranda byly publikovány ve čtyřech časopiseckých článcích, ve sbornících čtyř mezinárodních a dvou národních konferencí – lze tedy konstatovat, že výsledky jsou kvalitní a že s jádrem původního výzkumu doktoranda byla odborná veřejnost dostatečně seznámena.

Ing. Martin Dostal v dizertační práci prokázal schopnost výzkumné práce. S ohledem na výše uvedené *doporučuji dizertační práci k obhajobě a navrhuji udělení akademického titulu „philosophiae doctor"* jejímu autorovi.

V Ostravě dne 12. března 2015

doc. RNDr. Petr Šaloun, Ph.D.
katedra informatiky
FEI VŠB-Technická univerzita Ostrava
17. listopadu 15
708 33 Ostrava-Poruba
e-mail: petr.saloun@vsb.cz