POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	**Implementation of a learning to rank system**
Jméno autora:	**Bc. Tomáš Trnka**
Typ práce:	diplomová
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	Katedra počítačů
Oponent práce:	Ing. Tomáš Tunys
Pracoviště oponenta práce:	Katedra kybernetiky

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání **průměrně náročné**

Hodnocení náročnosti zadání závěrečné práce.

Náročnost zadání je zcela v souladu s typem závěrečné práce a neshledávám jej více než průměrně náročné.

Splnění zadání **splněno**

Posuďte, zda předložená závěrečná práce splňuje zadání. V komentáři případně uveďte body zadání, které nebyly zcela splněny, nebo zda je práce oproti zadání rozšířena. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.

Zadání bylo splněno v plném rozsahu včetně nepovinné části (bod zadání č. 6)

Zvolený postup řešení **správný**

Posuďte, zda student zvolil správný postup nebo metody řešení.

Zvolený postup shledávám za správný. Jednotlivé části práce, ale hlavně část rešeršní a experimentální si zasloužily větší pozornost. K jednotlivým výhradám se odvolávám níže.

Odborná úroveň **C - dobře**

Posuďte úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.

Odbornou úroveň předložené práce musím klasifikovat jako dobrou. Práce obsahuje stručný popis modelů pro řazení, které nejsou řádně odcitovány, a jejich kategorizaci, která je chybná (viz. část Výběr zdrojů …). Modely jsou popsány tak stručně, že bez jejich předchozí znalosti nemá čtenář možnost pochytit, jak vlastně fungují. Celá kapitola o řadících modelech mlčky přeskakuje jeden důležitý bod, jímž je obtížnost problematiky řazení, která dala vzniknout rozmanité řadě modelů, které jsou specifické svým přístupem. Velké chyby se řešitel dopouští při popisu LambdaRanku jako pouhého zrychlení RankNetu. Nešlo o pouhé zrychlení učení, jak je v práci uvedeno! Tyto nedostatky dle mého názoru vycházejí ze špatné práce se zdroji, jelikož algoritmy „learning to rank“ nejsou ve výuce (pokud dobře vím) pokryty. V popisu 2 metrik řazení mi chybělo zdůvodnění, proč byla nakonec zvolena metrika ERR pro budoucí experimenty. Opravdu matoucí je popis formátu dat a jejich zpracování – např. popis inverzního indexu neobsahuje specifikaci toho, co vše obsahuje. Co se týče zpracování obsahu dokumentů, z textu vyplývá, že byla použita normalizace (vymazání diakritiky, apod.) poté stemming a (možná zároveň) lematizace, což není logicky správný postup. Dále není vůbec jasné, jaké stupně relevance byly jednotlivým dokumentům přiřazeny, tedy nikde není uvedena škála, se kterou se pracuje. Při převodu „kliků“ z logů na relevance pro offline učení řadících algoritmů se řešitel uchýlil k ad-hoc řešení bez toho, aby se pokusil najít a prostudovat relevantí literaturu, které je v tomto ohledu opravdu mnoho. Nejhůře z hlediska odbornosti a chudě (z hlediska obsahu) vypadá experimentální část práce, což je vzhledem k oboru řešitele překvapivé. Experimenty spočívaly v opakovaném pouštění algoritmů z knihovny RankLib, při nichž bylo monitorováno, jaký vliv má rozšiřování sady příznaků na výslednou utilitu (ERR@20). Nikde není popsáno s jakou sadou parametrů byly algoritmy poušteny, či kolik parametrizací modelů bylo cross-validováno (z obrázků vyvozuje 10 per model). U cross-validace není uvedeno kolik foldů tvoří trénovací, kolik validační a ko-

lik testovací sada. Dále není nikde popsáno, jakým způsobem byly dotazy rozděleny mezi jednotlivé foldy. Kompletně chybí hlubší analýza přínosu příznaků z jednotlivých skupin (ručně udělané, w2v, dbp), která se v případě regresních stromů, které ve výsledku byly vybrány jako nejlepší, nedělá obtížně, a naopak může přinést velice cenné informace pro budoucí práci (výběr příznaků). Dále mi úplně chybělo reportování učících křivek u nejlepší parametrizace každého z modelů – není např. vůbec jasné zda u komplexnějších (state-of-art) modelů nedochází k přeučování, či naopak u jednodušších k ne-doučování – shrnuto, jak vypadá rychlost učení z hlediska konvergence u jednotlivých modelů. Dle mého názoru se šlo veškerým výšše uvedeným výhradám vyhnout, kdyby bylo příslušným částem věnováno více péče a času, kterého si nemyslím, že muselo by nad rámec času vyhrazeného pro diplomovou práci. Neuvádím podle mého názoru nic, k čemu by se řešitel v průběhu vypracovávání práce nedostat (ať přímo či nepřímo), ale do práce to nezanesl.

Formální a jazyková úroveň, rozsah práce **A - výborně**

Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.

Vzhledem k tomu, že řešitel využil poskytnutou šablonu, je typografická stránka práce na vysoké úrovni. Jedinou výtkou ke stylistice je odkazování kapitol, obrázků a tabulek – většinou se vyskytují velkým písmenem, ale v práci je použito (ale konzistentě) malé (např. Figure 3. 4). Práce je psána velice dobrou angličtinou. Většinu chyb tvoří překlepy, ale těch je na celou práci zanedbatelně málo a neruší plynulost jejího čtení. Výčet několika chyb nevztahujících se nutně k textu: rozbitý vzorec v definici Cosine Similarity, obrázek v popisu příznaku Text Length má být 3.11 namísto 3.12, obrázky se statistikami nemají z pravidla popsané osy, chybí ukončovací závorka v definici ztrátové funkce RankNetu. Nakonec opravdu nerozumím, proč byly obrázky s výsledky experimentů rozděleny mezi kapitolu 5 a Apendix B, čtenář musí takto zbytečně listovat, aby dohledal příslušný obrázek a je těžké mezi nimi poté hledat rozdíly.

Výběr zdrojů, korektnost citací **C - dobře**

Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení závěrečné práce. Charakterizujte výběr pramenů. Posuďte, zda student využil všechny relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

Korektnost citací a jejich počet tvoří jeden z mnoha nedostatků diplomové práce. Počet citací tak důležitý není, pokud odcitované zdroje dostatečně pokrývají popisovanou problematiku, což v případě této práce říci nelze. Opravdu nerozumím, proč řešitel nevěnoval větší pozornost správnému odcitování většiny konceptů a algoritmů uvedených v kapitole 2 (rešerše řadících algoritmů a metrik) – celá tato kapitola obsahuje pouhé 6 citací. Kapitola obsahuje 2 řádně odkázané definice metrik ale za to popis 9 modelů řazení, z nichž pouze 4 jsou přímo odkázány (RankBoost, ListNet, Coordinate Ascent a LambdaMART). Popisy téměř všech modelů jsou tak stručné, že bez odkazu na zdroj nemá čtenář velkou šanci pochopit, na čem jsou založeny a jak fungují. Není se poté možná čemu divit, že se řešitel dopustí zásadních chyb jako např. členění modelů pouze do 2 kategorií – **pairwise** a **listwise**. Na jednu (dle mého názoru důležitou) kategorii se zde zapomnělo, a to **pointwise**. Tuto kategorii tvoří „klasické“ klasifikační/regresní modely strojového učení, které řeší úlohu řazení převedenou na úlohu kategorické klasifikace či regrese – mezi tyto modely patří např. MART či Random Forest, které řešitel chybně zařadil do kategorie **listwise**. Shodou okolností tyto „triviální“ modely fungují v praxi na srovnatelné úrovni se state of the art modelem, jímž je (diskutabilně) LambdaMART. Jako poznámku uvádím, že specifikace knihovny RankLibu, na které řešitel postavil své řešení, obsahuje pro každý z popisovaných algoritmů vlastní odcitovaný zdroj (viz.: https://people.cs.umass.edu/~vdang/ranklib.html), kdyby ji řešitel využil, mohl rozšířit nejen počet citací, ale i své znalosti ohledně těchto modelů.

Další komentáře a hodnocení

Vyjádřete se k úrovni dosažených hlavních výsledků závěrečné práce, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, publikačním výstupům, experimentální zručnosti apod.

Vložte komentář (nepovinné hodnocení).

III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE

Shrňte aspekty závěrečné práce, které nejvíce ovlivnily Vaše celkové hodnocení. Uveďte případné otázky, které by měl student zodpovědět při obhajobě závěrečné práce před komisí.

Negativně se do mého hodnocení promítla úroveň rešeršní a experimentální části práce. Řešitel určitě musel udělat hromadu práce a k dobru určitě beru v potaz, že se řešitel musel potýkat s aspekty reálného návrhu funkčního softwarového řešení, které bylo nutné nasadit on-premise, ale vzhledem k jeho specializaci, bohužel, nemůžu hodnotit práci lepší známkou než C - dobře.

Mé hodnocení je, bohužel, vesměs kritické, ale práce má také světlé stránky a obsahuje řadu dobrých částí – líbil se mi vlastní návrh příznaků, obzvlášť z DBpedie, a všemi oblíbené zapojování W2V do reprezentací dokumentů. Software je také napsán čistě a čitelně. Shrnuto, tak že k tomu, aby byla práce výborná ji chtělo opravdu věnovat jen větší péči.

Otázky:

- V jakém předstihu byl řešitel/vedoucí v kontaktu s průmyslovým partnerem, od kterého získal data? Kdo rozhodoval/zvolil, co se bude logovat a v jakém formátu?
- Na základě jakého kritéria byly algoritmy v rešeršní části kategorizovány (pairwise/listwise)? Patří MART či RandomForest dle tohoto kritéria do listwise přístupů?
- Co skutečně přispělo k vyšší efektivitě (hlavně z hlediska řadících metrik) LambdaRanku v porovnání s jeho předchůdcem RankNetem? (Nápověda: viz. reference [1])
- Můžete popsat rozdíl mezi MARTem a Random Forestem? Co má vliv na rychlost učení těchto modelů, v čem může být výhoda jednoho modelu nad druhým?
- Vzhledem k nejlepším modelům – MART/RandomForest – můžete říci něco více o vlivu jednotlivých příznaků na výslednou chybu natrénovaných modelů? Analýza příznaků byla provedena pouze na globální úrovni.
- LambdaMART model bývá označován za state-of-art model řazení, ale z výsledků práce vyšel jako nejhorší? Můžete se pokusit vysvětlit, proč je tomu tak ve vašem případě?

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **C - dobře.**

Datum: 6. 6. 2016 Podpis: