



Hodnocení vedoucího závěrečné práce

Vedoucí práce: Ing. Martin Labský, Ph.D.
Student: Bc. Adriána Majtánová
Název práce: Aplikace Learning to Rank metod pro malé datasety
Obor / specializace: Znalostní inženýrství
Vytvořeno dne: 7. června 2023

Hodnotící kritéria

1. Splnění zadání

- ▶ [1] zadání splněno
- [2] zadání splněno s menšími výhradami
- [3] zadání splněno s většími výhradami
- [4] zadání nesplněno

Práce se zabývá praktickým tématem zlepšování relevance vyhledávání v textových dokumentech v případě, že máme k dispozici určité množství trénovacích dat. Toto je zajímavé téma mj. pro vyhledávání ve vnitrofiremních dokumentech, kde jsou často v provozu Information Retrieval (IR) systémy, jejichž uživatelé dávají zpětnou vazbu k relevanci nalezených dokumentů např. "proklikem" nebo jejich explicitním hodnocením, jako je klikání na "palec nahoru/dolů" ikonky. Toto typicky generuje malé až střední množství "trénovacích dat" v podobě historických dotazů a dokumentů s jejich hodnocenou relevancí k danému dotazu.

Zadání práce si klade za cíl jednak zmapovat současné možnosti využití metod Learning to Rank (LTR) pro zlepšení relevance výsledků, a v praktické části pak definuje 3 cíle a s nimi spojené otázky: (1) experimentálně ověřit, zda aplikace metod LTR nad rámec klasického IR systému vede ke zlepšení přesnosti vyhledávání, (2) kvantifikovat, zda a jakého zlepšení lze dosáhnout i pomocí malých objemů trénovacích dat, a (3) zda aplikace LTR dokáže vylepšit i IR systém, který využívá metod sémantického vyhledávání pomocí neuronových sítí typu transformer.

Konstatuji, že:

- práce explicitně a srozumitelně definuje výše uvedené cíle,
- obsahuje čtivý popis "state of the art" v oblasti IR a aplikace LTR pro IR,
- autorka získala data, implementovala SW a provedla experimenty, které umožňují odpovědět na všechny 3 experimentální cíle,
- práce odpovídá na všechny 3 experimentální cíle a obsahuje analýzu výsledků,
- autorce se podařilo na anglickém i českém datasetu prokázat zlepšení relevance nalezených dokumentů pomocí LTR (cíl 1),

- toto zlepšení dále kvantifikuje pro několik velikostí trénovacích dat (cíl 2) i pro zlepšení IR systému, který již v základu využívá sémantické vyhledávání (cíl 3).

2. Písemná část práce

92/100 (A)

Rozsah práce je přiměřený cílům zadání a rozsahu doprovodného kódu a experimentů. Autorka správně cituje relevantní zdroje.

Oceňuji, že práce je napsána v anglickém jazyce, což ji zpřístupňuje většímu počtu čtenářů.

Práce má logickou strukturu, dobře se v ní orientuje.

Kapitoly 1 a 2 poskytují úvod do problematiky vyhledávání v textových dokumentech, popisují základní IR metody a přístupy k LTR.

Kapitola 2 obsahuje systematické členění LTR metod podle reprezentace vstupních příkladů a vysvětluje principy LTR algoritmů.

Uvedené informace jsou relevantní a korektní.

Další kapitoly 3-5 zabírají cca 2/3 práce a popisují vlastní přínos autorky, konkrétně výběr a popis datových sad,

popis provedených experimentů a jejich implementaci, a shrnují dosažené výsledky v kontextu otázek položených v zadání práce.

V závěru autorka shrnuje odpovědi na vytýčené otázky, diskutuje využitelnost výsledků pro praxi, a zmiňuje možné budoucí kroky.

Práce obsahuje drobné formátovací nebo jazykové nedostatky, které ale neubírají práci na kvalitě:

- Z hlediska anglického jazyka obsahuje práce drobné stylistické prohřešky jako je občasná nadužívání členů, ale toto je velmi přijatelná cena za to, že je práce přímo čitelná pro anglicky mluvící čtenáře.

- Drobná nekonzistence ve vkládání prázdných listů oddělujících kapitoly - před kapitolami 3, 4, a 5 je vždy jeden prázdný list, zatímco před kapitolou 2 chybí.

- Při prezentaci výsledků v tabulkách v kapitole 5 by bylo dále dobré uvést v těchto tabulkách i relativní zlepšení metrik v následujících řádcích, tj. např. že relativní zlepšení MRR nebo přesnosti@10 je oproti základnímu systému 7 procent.

3. Nepísemná část, přílohy

94/100 (A)

Přílohy zahrnují:

- vlastní kód napsaný pro účely zpracování datových sad popsanych v kapitole 3 a běh experimentů popsanych v kapitole 4

- kód je možné spustit a zreprodukovat tak dosažené výsledky reportované v kapitole 5

Kompletní datové sady nejsou z důvodů rozsahu a publikovatelnosti součástí příloh práce, nicméně autorka přibalila menší část MS Marco dat.

Pro účely replikace celých experimentů je možné celý dataset MS Marco stáhnout jako volně dostupný dataset z odkazů zmíněných v práci, zatímco pro DareCzech je třeba napsat týmu Seznam.cz email s žádostí o tato data.

Co se týče kvality a zpracování SW:

- autorka využila Python jako dobrou praktickou volbu vzhledem k dostupnosti knihoven

- samostatně vyhledala a použila dostupné vhodné knihovny (např. ir-datasets pro práci s datovými sadami pro IR nebo hnsplib pro sémantické vyhledávání ve velké množině vektorů)

- pro snazší reprodukci experimentů je připojen jak celý Python projekt, tak i Python notebook, který umožňuje snazší reprodukci části experimentů
- postup pro spuštění experimentů je srozumitelně popsán krok za krokem v "readme" souboru

Několik námětů na drobná formální vylepšení ke zdrojovým souborům:

- důslednější specifikace type hints i pro návratové hodnoty funkcí, někde by toto pomohlo čitelnosti
- výběr a přidání licence jako např. MIT licence

4. Hodnocení výsledků, jejich využitelnost

96/100 (A)

Zadání práce bylo motivováno praktickou potřebou zjistit, zda jsou LTR metody vhodné pro zlepšení přesnosti firemních IR systémů v momentě, kdy začnou být vlivem provozu dostupná trénovací data v podobě prokliků nebo explicitního hodnocení relevance nalezených výsledků.

Práce podle očekávání v souladu s literaturou potvrdila, že tradiční vyhledávací systémy založené na TF-IDF mohou být vhodně doplněny sémantickým vyhledáváním, které vykazuje velké zisky pro obě testované datové sady a metriky.

Hlavními relativně novými přínosy této práce ale jsou:

- Zjištění, že i od jednotek tisíc trénovacích dotazů lze úspěšně aplikovat LTR a očekávat zlepšení relevance, a demonstrace této skutečnosti na anglickém i českém datasetu.
- Zjištění, že i v případě "chytrého" vyhledávacího systému v základu využívajícího sémantickou podobnost může LTR stále výrazně vylepšit relevanci nalezených dokumentů, což je opět demonstrováno na "malých" anglických i českých datech.

Autorka výše uvedené poznatky demonstruje zlepšením MRR (mean reciprocal rank) pro různé velikosti trénovacích množiny na datasetu MS Marco.

Slibné jsou zejména výsledky pro trénovací velikosti označené jako "mini" (3000 dotazů) a "small" (7000 dotazů), kde je demonstrováno relativní zlepšení MRR o cca 7-8% oproti základnímu systému bez LTR (Tabulka 5.1).

Pokud je základní systém "chytřejší" a využívá sémantické vyhledávání, je naměřený přínos LTR nižší - v řádu relativních 2% (Tabulka 5.2) - což je ale stále zajímavé zlepšení.

Pro česká DareCzech data je pro trénovací sadu označnou jako "small" (necelých 2000 dotazů) pozorováno výraznější zlepšení okolo 10-12% relativně pro metriku Přesnost@10 (Tabulky 5.3-5.4) a to i v případě sémantického vyhledávání (Tabulky 5.5-5.6).

Experimenty provedené v práci dobře doplňují LTR články jako je Shuguang Han, 2020, převážně se zabývající aplikací LTR na větší množství trénovacích dat.

Práce tak ukazuje praktickou dobře zreplicovatelnou metodu, jak lze zlepšit tradiční nejen vnitrofiremní vyhledávače.

5. Aktivita studenta

- [1] výborná aktivita
- [2] velmi dobrá aktivita
- [3] průměrná aktivita
- [4] slabší, ale ještě dostatečná aktivita
- [5] nedostatečná aktivita

Autorka práce byla během tvorby práce iniciativní zejména v rámci vlastního vlastního přínosu práce, tj. experimentů, napsání a ladění potřebného kódu a opatření vhodných datových sad.

Postup práce jsme pravidelně konzultovali.

Pro účely práce získala od výzkumného týmu Seznam.cz dataset DareCzech, který vedle MS Marco použila na zodpovězení vytýčených otázek.

Nad rámec těchto dvou v práci popsaných datasetů experimentovala s několika dalšími.

Nutno také ocenit poměrně náročnou evaluaci kompletních MS Marco dat na GPU HW, který měla pro práci k dispozici.

6. Samostatnost studenta

- ▶ [1] **výborná samostatnost**
- [2] velmi dobrá samostatnost
- [3] průměrná samostatnost
- [4] slabší, ale ještě dostatečná samostatnost
- [5] nedostatečná samostatnost

Autorka vykazovala po celou dobu tvorby práce vysokou iniciativu a samostatnost jak v rešeršní části práce, tak v průběhu řešení experimentů, vývoje potřebného SW, i jeho běhu na vhodném HW.

Celkové hodnocení

95 /100 (A)

Celkově hodnotím práci jako velmi povedenou, a to z těchto důvodů:

- vysoká iniciativa autorky při rešerši, výběru datových sad a použitých LTR metod.
- velký rozsah provedených experimentů se zajímavými částečně novými výsledky na datech v českém i anglickém jazyce,
- výsledky prakticky využitelné pro zlepšení vyhledávání nejen ve vnitrofiremních dokumentových systémech,
- srozumitelně napsaná práce v anglickém jazyce, což ji činí přístupnou širšímu kruhu čtenářů.

Instrukce

Splnění zadání

Posudte, zda předložená ZP dostatečně a v souladu se zadáním obsahově vymezuje cíle, správně je formuluje a v dostatečné kvalitě naplňuje. V komentáři uveďte body zadání, které nebyly splněny, posudte závažnost, dopady a případně i příčiny jednotlivých nedostatků. Pokud zadání svou náročností vybočuje ze standardů pro daný typ práce nebo student případně vypracoval ZP nad rámec zadání, popište, jak se to projevilo na požadované kvalitě splnění zadání a jakým způsobem toto ovlivnilo výsledné hodnocení.

Písemná část práce

Zhodnoťte přiměřenost rozsahu předložené ZP vzhledem k obsahu, tj. zda všechny části ZP jsou informačně bohaté a ZP neobsahuje zbytečné části. Dále posudte, zda předložená ZP je po věcné stránce v pořádku, případně vyskytují-li se v práci věcné chyby nebo nepřesnosti.

Zhodnoťte dále logickou strukturu ZP, návaznosti jednotlivých kapitol a pochopitelnost textu pro čtenáře. Posudte správnost používání formálních zápisů obsažených v práci. Posudte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 52/2021, článek 3.

Posudte, zda student využil a správně citoval relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami. Zhodnoťte, zda převzatý software a jiná autorská díla, byly v ZP použity v souladu s licenčními podmínkami.

Nepísemná část, přílohy

Dle charakteru práce se případně vyjádřete k nepísemné části ZP. Například: SW dílo – kvalita vytvořeného programu a vhodnost a přiměřenost technologií, které byly využité od vývoje až po nasazení. HW – funkční vzorek – použité technologie a nástroje, Výzkumná a experimentální práce – opakovatelnost experimentů.

Hodnocení výsledků, jejich využitelnost

Dle charakteru práce zhodnoťte možnosti nasazení výsledků práce v praxi nebo uveďte, zda výsledky ZP rozšiřují již publikované známé výsledky nebo přinášející zcela nové poznatky.

Aktivita studenta

V souvislosti s průběhem a výsledkem práce posudte, zda byl student během řešení aktivní, zda dodržoval dohodnuté termíny, jestli své řešení průběžně konzultoval a zda byl na konzultace dostatečně připraven.

Samostatnost studenta

V souvislosti s průběhem a výsledkem práce posudte schopnost studenta samostatně tvůrčí práce.

Celkové hodnocení

Shrňte stránky ZP, které nejvíce ovlivnily Vaše celkové hodnocení. Celkové hodnocení nemusí být aritmetickým průměrem či jinou hodnotou vypočtenou z hodnocení v předchozích jednotlivých kritériích. Obecně platí, že bezvadně splněné zadání je hodnoceno klasifikačním stupněm A.