FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Posudek oponenta závěrečné práce

Oponent práce:	Ing. Karel Klouda, Ph.D.
Student:	Bc. Petr Kasalický
Název práce:	Porovnání online a offline evaluačních metrik v doporučovacích systémech
Obor / specializace:	Znalostní inženýrství
Vytvořeno dne:	2. června 2021

Hodnotící kritéria

1. Splnění zadání

▸ [1] **zadání splněno**
[2] zadání splněno s menšími výhradami
[3] zadání splněno s většími výhradami
[4] zadání nesplněno

Všechny body zadání byly splněny. Jedná se přitom o dosti náročné zadání, ať už pracností (práce nejen s reálnými a velikými daty ale i s produkčně nasazenými systémy) tak i svou nezvyklostí, která je důsledkem málo probádaného typu experimentu.

2. Písemná část práce 98 /100 (A)

Práce je psána velmi dobrou a srozumitelnou angličtinou. Je velice dobře strukturována a je velmi dobře zvoleno značení. Díky tomu je poměrně jednoduché pochopit, co je popisováno a jak přesně byl nastaven a vyhodnocován experiment a příslušné metriky. Přitom svou komplexností a vysokou mírou zaměření na detail (drobné nuance v nastavení např. měření recallu) by se mohlo snadno stát, že popis experimentu bude těžko srozumitelný.
V práci jsem našel minimální množství chyb jakéhokoli typu. Abych alespoň něco zmínil: Občas se mi nezdála volba interpunkce a občas některé zkratky a pojmy nebyly úplně přehledně (nebo vůbec) zavedeny (např. rel_j a REL_P na straně 17, obecně tento odstavec o DCGa nDCG považuji za nepatřičný - dále se o těchto metrikách nemluví).

3. Nepísemná část, přílohy 98 /100 (A)

Nepísemná část práce má formu několika poměrně přehledně popsaných skriptů v jazyce Python. V rámci experimentu byly tyto skripty napojeny na produkční prostředí firmy Recombee, které samozřejmě přiloženo není a je tak i nemožné experiment zopakovat.

To je ale přirozený důsledek situace a také zároveň důvod jedinečnosti tohoto experimentu.

4. Hodnocení výsledků, jejich využitelnost 100 /100 (A)

Hlavním výsledkem práce je poctivě a přesvědčivě provedený experiment, který ukazuje, že jeden ze základních pilířů budování a zkoumání doporučovacích systémů je vratký: Tedy že hojně používaná trénovací offline metrika recall ne vždy koreluje se skutečnou mírou toho, jak je doporučování úspěšné (v práci měřeno pomocí CTR) v reálném provozu. Vedlejším výsledkem je pak poměrně silně podpořené tvrzení, že nově představená metodika měření recallu (vzniklá kombinací dříve představených metod) je přeci jen lepší volbou z pohledu korelace s CTR než jiné.

Celkové hodnocení 99 /100 (A)

Jedná se o skvělou práci, která svému oboru přináší opravdu užitečná zjištění. Od klasického výzkumu se navíc liší tím, že zkoumané modely byly konfrontovány s reálným nasazením a reálnými uživateli, což z práce dělá opravdu nebývalé dílo. Vzhledem k tomu a také vzhledem ke všemu výše napsanému navrhuji práci hodnotit známkou A jako výbornou.

Otázky k obhajobě

1) Při aktualizaci ALS embeddingů (přidání nových dat, interakcí atd.) se počítá ALS kompletně znovu, nebo jsou známá nějaká "updatovací pravidla", která dovolují pouze aktualizovat nové výpočty?

2) Mohl byste přesněji vysvětlit, na čem přesně je měřen recall při updatování modelů o nové uživatele? Přesněji: Jak se projeví různé parametery t__i ve vzorci (2.5) na straně 29 a jaký je jejich vztah k výpočtu a množině testovacích uživatelů.

3) V jakých jednotkách je měřen parametr d (str. 38, kde se píše, že je to 10)?

4) Jak moc je problematické daný experiment zveřejnit spolu se zkoumanými (anonymizovanými) daty?

Instrukce

Splnění zadání

Posuďte, 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, posuďte 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 posuďte, 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. Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 26/2017, článek 3.

Posuďte, 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.

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.