Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-0000-72397

Bc. Martin Šrank

Podpora diverzity a aktuálnosti informačných bulletinov v systéme pre odpovedanie na otázky

Diplomová práca

Študijný program: Informačné systémy
Študijný odbor: 9.2.6 Informačné systémy
Miesto vypracovania: Ústav informatiky, informačných systémov a softvérového inžinierstva, FIIT STU, Bratislava
Vedúci práce: Ing. Ivan Srba, PhD.

Máj 2017

Anotácia

Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ
Študijný program: Informačné systémy
Autor: Bc. Martin Šrank
Diplomová práca: Podpora diverzity a aktuálnosti informačných bulletinov v systéme pre odpovedanie na otázky
Vedúci práce: Ing. Ivan Srba, PhD.
Máj 2017

Informačné bulletiny predstavujú štandardný spôsob ako informovať používateľov v online komunitách o novom alebo zaujímavom obsahu. Ich význam je ešte väčší v online komunitách ktoré produkujú veľké množstvo používateľmi vytvoreného obsahu, akými sú aj systémy pre odpovedanie na otázky. Napriek tomu mnohé populárne CQA systémy ponúkajú iba generické informačné bulletiny, ktoré nijakým spôsobom nereflektujú záujmy používateľa alebo diverzitu odporúčaného obsahu.

Cieľom našej práce je analyzovať existujúce prístupy k personalizovanému odporúčaniu v CQA systémoch a navrhnúť metódu automatického vytvárania personalizovaných informačných bulletinov pre jednotlivých používateľov. Zameriavame sa na zlepšenie diverzity a aktuálnosti odporúčaného obsahu ako spôsobu prevencie vzniku *filtračnej bubliny* a zvýšenie celkovej spokojnosti používateľov a ich interakcie s informačným bulletinom.

Annotation

Slovak University of Technology Bratislava
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES
Degree Course: Information Systems
Author: Bc. Martin Šrank
Master thesis: Improving Diversity and Freshness of Newsletters in Community Question Answering Systems
Supervisor: Ing. Ivan Srba, PhD.
May 2017

Newsletters represent a standard way to inform users of online communities about new or interesting content. Their importance is even greater in online communities producing large amounts of user-created data, such as Community Question Answering systems. Nevertheless, many popular CQA systems only offer generic newsletters, which do not take into account users' interests or diversity of the recommended content.

The aim of this work is to analyze existing approaches in personalized content recommendation in CQA systems and design a method for automatic creation of personalized newsletters for individual users of CQA systems. We want to focus on improving the diversity and freshness of the recommended content as a way to prevent *filter bubbles* and improve overall user satisfaction and engagement with the newsletter.

Obsah

1 Úvod

Informačné bulletiny (angl. *newsletters*) sú stále jednou z najčastejšie používaných foriem informovania používateľov webových portálov o novinkách, akciách alebo zaujímavom obsahu na webe. Používatelia radi využívajú informačné bulletiny svojich obľúbených webových portálov, pretože predstavujú jednoduchý a prehľadný spôsob prezentácie obsahu, ktorý je navyše doručený pohodlne priamo do používateľovej e-mailovej schránky.

Informačné bulletiny zastávajú ešte väčšiu rolu v rámci online komunít, ktoré produkujú veľké množstvo používateľmi vytváraného obsahu. Medzi populárne druhy takýchto online komunít patria aj systémy pre odpovedanie na otázky (angl. *Community Question Answering systems – CQA*). Veľké množstvo obsahu, ktoré v týchto systémoch vzniká, si vyžaduje nasadzovanie personalizačných techník za účelom poskytnutia relevantného obsahu používateľom.

Výskum v oblasti CQA systémov sa v súčasnosti skôr orientuje na skúmanie správania používateľov, kladenia otázok a odpovedania. Problematike informačných bulletinov v doméne CQA systémov zatiaľ nebola venovaná dostatočná pozornosť.

Cieľom našej práce je analyzovať súčasný stav výskumu odporúčania otázok v doméne CQA systémov a navrhnúť riešenie pre tvorbu personalizovaných informačných bulletinov v systémoch pre odpovedenaie na otázky. V našej práci sa chceme zamerať predovšetkým na skúmanie vplyvu zavedenia diverzity a podpory aktuálnosti na úspešnosť personalizovaného odporúčania otázok vo forme informačných bulletinov.

2 Informačné bulletiny

Informačné bulletiny sú aj v súčasnosti jedným z najrozšírenejších spôsobov, ako v online prostredí informovať používateľov o dianí na webovom portáli. Prevádzkovatelia webových portálov využívajú informačné bulletiny na predstavenie nového obsahu, akciového tovaru, zaujímavostí z určitej oblasti alebo špeciálnych ponúk pre svojich používateľov a zákazníkov. Informačné bulletiny sú tiež využívané ako prostriedok pre motiváciu používateľov k opätovnej návšteve webového portálu.

Informačné bulletiny spravidla nadobúdajú formu e-mailu, ktorý je zvyčajne v pravidelných intervaloch doručovaný do schránok používateľov, ktorí o jeho doručovanie prejavili záujem.

2.1 Systémy využívajúce informačné bulletiny

Informačné bulletiny sú efektívnym spôsobom dosahu na používateľov vo viacerých druhoch webových portálov, pričom pre každú kategóriu portálov spĺňajú mierne odlišné ciele, čomu sa prispôsobuje aj ich obsah:

- **Marketing / internetové obchody**
  Najčastejšie informačné bulletiny rozposielajú svojim používateľom práve internetové obchody, zľavové portály a iné marketingové weby. Tieto bulletiny zvyčajne ponúkajú používateľom zaujímavé produkty, rôzne zľavy alebo exkluzívne ponuky. Cieľom týchto bulletinov je získať pozornosť používateľa po tom, ako už úspešne na internetovom obchode nakúpil, aby tak urobil znova.
- **Tématické weby/blogy**
  Webové stránky alebo blogy, ktoré produkujú tématický obsah spravidla určený pre konkrétne záujmové skupiny používateľov, zvyčajne využívajú informačné bulletiny na informovanie používateľov o novom zaujímavom obsahu na stránke. Vzhľadom na kvantitu nového obsahu je zvyčajne prispôsobená aj frekvencia rozposielania bulletinov.
- **Komunitné portály**
  Komunitné portály sú odlišné hlavne tým, že väčšinu obsahu vytvárajú samotní používatelia. Informačné bulletiny týchto portálov spravidla obsahujú zoznam najnovších, najzaujímavejších, alebo najkontroverznejších príspevkov. Okrem toho môžu obsahovať aj informácie alebo správy od moderátorov a prevádzkovateľov portálu, prípadne obsah z iných pridružených webov, napr. blogov

- **CQA systémy**
  CQA systémy patria medzi komunitné portály, preto aj ich informačné bulletiny zdieľajú podobný cieľ a štruktúru. Odlišujú sa však tým, že zvyčajne tieto systémy produkujú veľmi veľa obsahu. Pre používateľa preto môže byť problematické nájsť zaujímavý alebo relevantný obsah. Práve tento problém by mali riešť informačné bulletiny.

Spôsob vytvárania informačných bulletinov nie je závislý len od druhu webového portálu, ale aj od množstva obsahu, ktorý tento portál produjuje.

Ručne zostavované bulletiny
Zostavovanie informačných bulletinov ručne autorom, správcom či moderátorom webového portálu je únosné len v prípade, že webový portál vyprodukuje za dané obdobie iba relatívne malé množstvo obsahu. Využívať sa môže hlavne v prípade tématických blogov, kde úzke zameranie cieľovej skupiny používateľov zároveň umožňuje autorovi zostaviť pomerne relevantný informačný bulletin, ktorý je pre používateľov zaujímavý a prínosný.

Automaticky generované generické bulletiny
Najčastejším druhom informačných bulletinov sú bulletiny, ktoré obsahujú len generický obsah, napr. zoznam najpredávanejších produktov v internetovom obchode, alebo produkty s najväčšími zľavami. Takýto bulletin sa zostavuje pomerne jednoducho, no jeho prínos je otázny vzhľadom na veľkú diverzitu obsahu aj používateľov.

Perosnalizované bulletiny
Najefektívnejším spôsobom, ako zostaviť relevantný bulletin pre veľké množstvo používateľov alebo z veľkého množstva obsahu, je využitie metód personalizácie a odporúčania napr. na základe predošlej aktivity používateľa. Takýto prístup umožňuje zostaviť pre každého používateľa bulletin, ktorý má najväčšiu šancu splniť svoj cieľ – či už je to nákup ďalších produktov, alebo zvýšenie návštevnosti portálu.

2.2 Problémy informačných bulletinov

Hlavným problémom informačných bulletinov je stále sa znižujúca miera interakcie používateľov s informačnými bulletinmi.

Štúdia spoločnosti Silverpop z roku 2012 [1] na vzorke informačných bulletinov 1124 spoločností ukázala, že počet používateľov, ktorí vôbec otvoria informačný bulletin sa pohybuje na úrovni 20% a stále klesá. Navyše konkrétne v oblasti technológií sa táto hodnota pohybuje len na 16,5%. Ešte menšia je miera preklikov (angl. *Click-through rate - CTR*), ktorá sa celkovo pohybuje na úrovni 5,4% a v prípade technologicky zameraných informačných bulletinov len

3,6%. Napriek tomu sa miera odhlásení z odoberania (angl. *unsubscribe rate*) pohybuje len na úrovni 2%.

Dôvodov, prečo používatelia prejavujú iba malý záujem o informačné bulletiny, ktoré im sú doručované, môže byť niekoľko. Jedným z takýchto dôvodov môže byť vysoká saturácia – používateľom chodí priveľké množstvo informačných bulletinov, dôsledkom čoho používatelia rezignujú a tieto e-maily ani neotvárajú. Hlavným nedostatkom informačných bulletinov, a zároveň dôvodom, prečo iba 5% používateľov klikne na obsah v informačnom bulletine, je však relevancia ponúkaného obsahu.

2.3 Relevancia v informačných bulletinoch

Množstvo webových portálov doručuje všetkým svojim používateľom presne ten istý obsah informačného bulletinu. Často je tento obsah vytváraný manuálne editormi, a zameriava sa len všeobecne na aktuálne dianie na danom webovom portáli. Takýto všeobecný informačný bulletin však nutne nemôže byť dostatočne relevantný pre značnú časť používateľov.

Riešením problému relevancie informačných bulletinov je vytváranie personalizovaných informačných bulletinov, ktoré každému používateľovi ponúkajú len ten obsah, ktorý je pre neho najzaujímavejší a najrelevantnejší.

2.4 Diskusia

Kvalitné informačné bulletiny sú obzvlášť dôležité pre webové portály, ktoré obsahujú veľké množstvo diverzného obsahu. Medzi takéto portály patria aj systémy pre odpovedanie na otázky – CQA systémy, ktoré tvorí primárne používateľmi vytváraný obsah. Pre tieto systémy nie je efektívne vytvárať informačné bulletiny ručne, ani poskytovať len generické informačné bulletiny.

3 CQA systémy

CQA systémy sú jednou z výrazných skupín webových portálov, ktoré sú založené na princípe používateľsky vytváraného obsahu. Tieto systémy umožňujú používateľom položiť otázky, ktoré nie je možné zodpovedať použitím štandardných vyhľadávačov [2] a zároveň odpovedať na otázky iných používateľov.

Napriek tomu, že väčšina CQA systémov sa spočiatku zameriava najmä na poskytnutie zmysluplnej odpovede na konkrétnu otázku, v súčasnosti je možné v prípade niektorých CQA systémov (napr. Stack Overflow) vnímať postupnú zmenu zamerania z jednorázových odpovedí na kolaboratívne vytváranie komplexnejších poznatkov s dlhodobou hodnotou [3]. Za týmto účelom CQA systémy implementujú hlasovanie a princíp reputácie ako spôsob podpory komunitného aspektu označovania najlepších odpovedí na položené otázky.

3.1 Druhy CQA systémov

CQA systémy možno kategorizovať do dvoch základných skupín podľa toho, na akú oblasť otázok sa tieto systémy zameriavajú.

Univerzálne CQA systémy
CQA systémy ako *Yahoo! Answers*, *Wiki Answers* alebo *Quora* nie sú zamerané na konkrétne oblasti a umožňujú používateľom pokladať otázky na akékoľvek témy [4].

Tento druh CQA systémov má štandardne vyšší počet používateľov aj aktivity ako úzko špecializované CQA systémy, no tiež tu existuje väčšia pravdepodobnosť výskytu nekvalitných, jednoduchých alebo neužitočných otázok a odpovedí, ako aj veľký počet duplicitných otázok, ktoré už boli zodpovedané. Zároveň sú univerzálne CQA systémy zamerané viac na samotný proces kladenia otázok a odpovedania na ne, než na vytváranie dlhodobo hodnotného obsahu.

Úzko špecializované CQA systémy
Opakom univerzálnych CQA systémov sú CQA systémy, ktoré sú špecializované na konkrétne oblasti záujmu. Medzi takéto CQA systémy patria napríklad jednotlivé komunity v rámci siete Stack Exchange, ktorá zahŕňa rôzne druhy komunít, od všeobecnejších, ako je napr. komunita venujúca sa matematike[1], po veľmi úzko špecializované, akými sú napr. komunity *Ask Ubuntu*[2]

[1] `https://math.stackexchange.com`
[2] `https://askubuntu.com`

alebo *Raspberry Pi*[3] venujúce sa konkrétnym produktom.

Tématicky zamerané CQA systémy majú väčší potenciál pre vznik dlhodobo hodnotného obsahu [3]. V rámci týchto systémov tiež vzniká množstvo prepojení medzi obsahom *(otázky podobného charakteru, riešenie problému v príbuznej oblasti)*, čo vedie k vzniku *znalostných sietí* (angl. *knowledge networks*) [5]. S cieľom zvýšiť hodnotu jednotlivých príspevkov tiež mnohé CQA systémy zavádzajú možnosť komunitnej úpravy otázok a odpovedí [6], čo vedie okrem zvýšenej aktivity aj k zvýšeniu vnímanej užitočnosti príspevku.

3.2 Problémy CQA systémov

CQA systémy sa musia vysporiadavať s tými istými druhmi problémov, ako iné kategórie systémov založené na používateľmi vytvorenom obsahu.

3.2.1 Problém dlhého chvosta v aktivite používateľov

Čím viac sa zvýrazňuje trend orientácie CQA systémov skôr na poskytovanie obsahu s dlhodobou hodnotou ako na samotné poskytnutie odpovede na položenú otázku, tým viac sa prehlbuje problém *dlhého chvosta* (angl. *long tail*). Ide o štandardný problém všetkých stránok zameriavajúcich sa na používateľmi vytváraný obsah, kedy je veľká väčšina používateľov týchto stránok len pasívnymi čitateľmi (angl. *lurkers*) a najväčšia časť obsahu je vytvorená len veľmi úzkou skupinou najaktívnejších používateľov.

V prípade CQA systému Stack Overflow sa podiel aktívnych používateľov (takých, ktorí za sledovaný mesiac pridali do systému aspoň jednu otázku alebo odpoveď) za marec 2017[4] pohyboval na úrovni 3% všetkých používateľov [7].

3.2.2 Variabilita v kvalite obsahu

Ďalším problémom CQA systémov je variabilná kvalita otázok a odpovedí v týchto systémoch. Zvyšujúcou sa popularitou CQA systémov narastá aj podiel obsahu s nízkou kvalitou, či už vo forme veľmi jednoduchých otázok alebo nedostatočne podrobných odpovedí, ako aj množstvo duplicitných otázok – otázok, ktoré už boli v systéme zodpovedané [7, 8].

[3] `https://raspberrypi.stackexchange.com`

[4] Výsledky za aktuálne obdobie boli získané prostredníctvom nástroja Stack Exchange Data Explorer – `https://data.stackexchange.com`

Jedným z riešení tohto problému, ktorý využíva aj CQA systém Stack Overflow, je komunitné zabezpečovanie kvality obsahu prostredníctvom moderátorov – používateľov s oprávnením upravovať, označiť duplikáty alebo vymazať obsah.

3.3 Informačné bulletiny v CQA systémoch

Význam informačných bulletinov narastá v rámci systémov pre odpovedanie na otázky, ktoré sú prominentným druhom online komunít produkujúcich veľké množstvo používateľmi vytváraného obsahu.

Súčasný výskum v oblasti CQA systémov [9] sa venuje predovšetkým oblastiam skúmania správania používateľov, smerovania a odporúčania otázok a kvality otázok a odpovedí v týchto systémoch. Problematike vytvárania informačných bulletinov v doméne CQA systémov zatiaľ nebola venovaná veľká pozornosť.

Mnohé populárne CQA systémy aj v súčasnosti ponúkajú svojim používateľom informačné bulletiny majúce iba generický charakter a nijakým spôsobom neuvažujú relevantnosť obsahu pre konkrétnych používateľov, prípadne informačné bulletiny neponúkajú vôbec.

3.3.1 Informačné bulletiny v sieti Stack Exchange

Sieť Stack Exchange[5], ktorá patrí medzi najpopulárnejšie CQA systémy súčasnosti, sa skladá z viac ako 160 samostatných komunít zameraných na rôzne oblasti. Stack Exchange ponúka používateľom všetkých komunít možnosť odoberať informačný bulletin, ktorý je doručovaný raz týždenne.

Informačné bulletiny komunít Stack Exchange obsahujú tri sekcie (Obr. 3.1). Prvá sekcia je rovnaká pre všetkých používateľov konkrétnej komunity a obsahuje zoznam najlepšie hodnotených nových otázok. Obsah nasledujúcich dvoch sekcií je náhodne generovaný. Tieto sekcie obsahujú najpopulárnejšie otázky z predchádzajúceho týždňa a náhodný výber nezodpovedaných otázok.

Používatelia CQA systému Stack Exchange nie sú s takýmto generickým informačným bulletinom spokojní [6]. Medzi problémy, ktoré najčastejšie používatelia vytýkajú súčasnému informačnému bulletinu patria:

- **Náhodne generovaný obsah** – Sekcia nezodpovedaných otázok obsahuje náhodný výber

[5] `https://stackexchange.com`

[6] `https://meta.stackexchange.com/q/247298`. Prevzaté 31.4.2017.

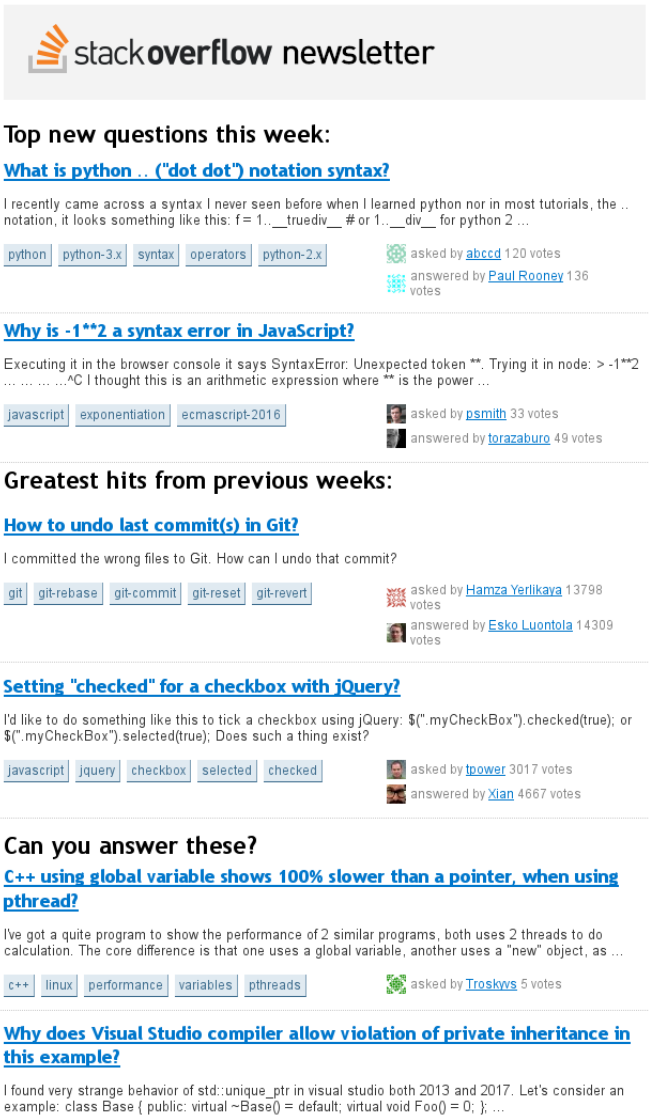
otázok bez odpovedí. Pravdepodobnosť, že používateľ vie na niektorú z nich odpovedať, je tak veľmi malá [7].

- **Absencia personalizácie** – Otázky v jednotlivých sekciách nijakým spôsobom nezohľadňujú používateľove obľúbené značky alebo jeho aktivitu. Dôsledkom hlavne pri väčších komunitách je tak nízka relevancia ponúkaného obsahu [8].
- **Malá rôznorodosť obsahu** – Hlavne pokročilejší a aktívnejší používatelia by chceli v informačnom bulletine vidieť okrem rôznych otázok aj iný obsah, okrem iného napr. rôzne štatistiky aktivity komunity, zoznam ocenených používateľov alebo príbuzný obsah z komunitných blogov [9].
- **Aktuálnosť obsahu** – Informačný bulletin niekedy obsahuje veľmi staré otázky, ktoré už nie sú relevantné[10].

Na všetky tieto problémy používatelia upozorňujú už dlhšiu dobu, tieto otázky majú pomerne veľkú podporu komunity, no napriek tomu žiaden z týchto problémov zatiaľ nebol adresovaný, a informačný bulletin ostáva aj naďalej generický a z veľkej časti plný náhodného obsahu.

Generický informačný bulletin stráca pre používateľov informačnú hodnotu, pretože najmä v prípade väčších komunít, akou je napríklad Stack Overflow[11], často obsahuje otázky, ktoré nie sú z oblastí záujmu používateľa.

[7] `https://meta.stackexchange.com/q/96758`
[8] `https://meta.stackexchange.com/q/110902`
[9] `https://meta.stackexchange.com/q/247298`
[10] `https://meta.stackoverflow.com/q/319095`
[11] `https://stackoverflow.com`



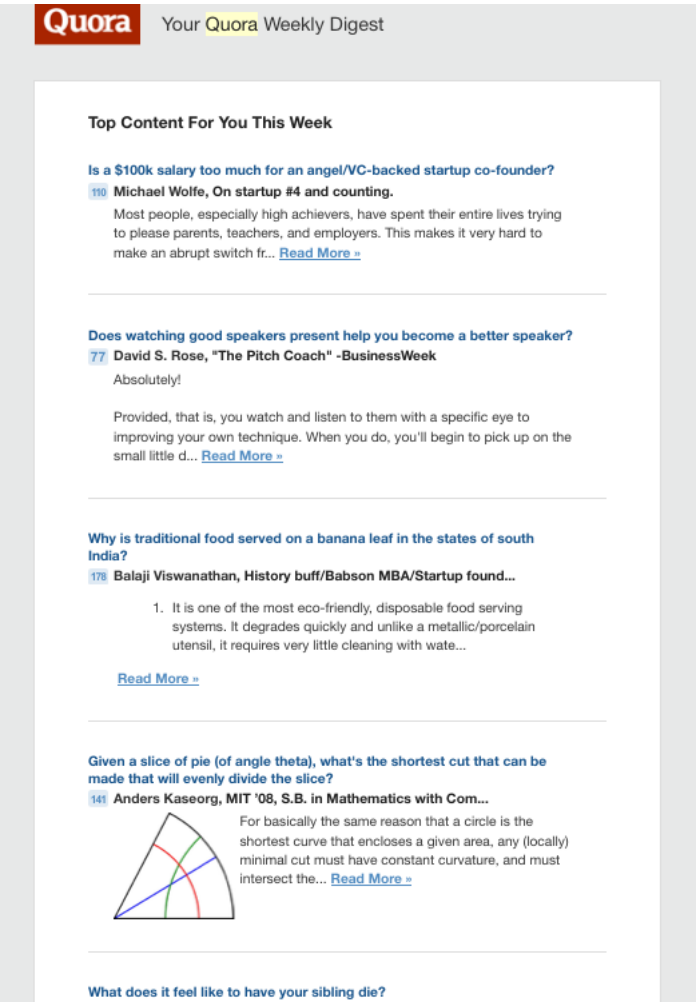
Obr. 3.1: *Informačný bulletin komunity Stack Overflow, 25. apríl 2017.*

3.3.2 Informačný bulletin portálu Quora

Quora[12] je CQA systém, ktorý nie je zameraný na konkrétnu oblasť záujmu, ale obsahuje otázky z rôznych tém. Quora ponúka svojim používateľom týždenný informačný bulletin (*Quora Weekly Digest*), ktorý obsahuje desať najzaujímavejších otázok za posledný týždeň a zoznam ľudí, ktorých používateľ potenciálne pozná.

Zoznam najzaujímavejších otázok pozostáva z editormi manuálne vybraného obsahu a algoritmicky vybraného obsahu, ktorý je personalizovaný pre každého používateľa zvlášť [13] (Obr. 3.2).

[12] `https://quora.com`
[13] `http://businessinsider.com/quora-emails-2012-8`



Obr. 3.2: *Informačný bulletin portálu Quora. Prevzaté 30.4.2017, [13]*

3.3.3 CQA systémy bez informačných bulletinov

Viaceré populárne CQA systémy svojim používateľom vôbec neponúkajú možnosť odoberať informačný bulletin. Medzi takéto systémy patrí napr. portál *Yahoo! Answers*[14], ktorý je určený na pokladanie otázok z akejkoľvek oblasti záujmu. Rovnako informačný newsletter neponúka ani ďalší všeobecne zameraný CQA systém – *Wiki Answers*[15]. CQA systém zameraný na podporu výučby *Askalot*[16] tiež v súčasnosti neponúka informačný newsletter, iba možnosť notifikácie používateľa prostredníctvom e-mailu o aktivite súvisiacej s jeho obsahom v rámci systému.

[14] `https://answers.yahoo.com`
[15] `https://answers.com`
[16] `https://askalot.fiit.stuba.sk`

3.4 Diskusia

CQA systémy patria medzi systémy s veľkým objemom používateľmi generovaného obsahu. Ako také sú vhodným kandidátom na implementáciu informačných bulletinov. Existujúce informačné bulletiny v skúmaných CQA systémoch, najmä v prípade platformy Stack Exchange, však majú veľké množstvo nedostatkov.

Riešením týchto problémov je vytváranie personalizovaných informačných bulletinov pre tieto systémy. Vhodným spôsobom na personalizáciu je využitie metód odporúčania, ktorým sa venuje kapitola 4.

4 Odporúčanie

4.1 Odporúčacie systémy

Odporúčacie systémy sú softvérové nástroje a techniky ktoré používateľom ponúkajú položky, ktoré by pre nich mohli byť nejakým spôsobom zaujímavé alebo užitočné [10]. Tieto odporúčania sú zvyčajne ponúkané za účelom pomôcť používateľovi rozhodnúť sa, aké články by si mal prečítať, alebo aký tovar si kúpiť.

Využívanie odporúčacích systémov je tiež pre používateľov vhodným spôsobom, ako zvládať problémy informačného zahltenia v dnešnom online svete. Ako také sa odporúčacie systémy stávajú jedným z najsilnejších a najpopulárnejších nástrojov v online komunitách.

Odporúčacie systémy typicky vytvárajú zoznam odporúčaní jedným z dvoch spôsobov – buď prostredníctvom *kolaboratívneho filtrovania* (angl. *Collaborative filtering*), alebo použitím *filtrovania založeného na obsahu* (angl. *Content-based filtering*) [11]. Tieto dva prístupy môžu byť tiež kombinované v hybridných odporúčacích systémoch.

4.1.1 Kolaboratívne filtrovanie

Odporúčacie systémy využívajúce kolaboratívne filtrovanie fungujú prostredníctvom získavania spätnej väzby používateľa vo forme hodnotení pre položky v danej doméne a využívajú podobnosti v hodnotení medzi viacerými používateľmi na určenie či určitý obsah odporučiť, alebo nie [11]. Metódy kolaboratívneho filtrovania možno ďalej rozdeliť na metódy založené na susednosti alebo na základe modelu.

Kolaboratívne filtrovanie na základe susednosti (angl. *Neighborhood-based Collaborative filtering*) vyberá skupinu používateľov podľa ich podobnosti k aktuálnemu používateľovi a použitím váženej kombinácie ich hodnotení vyberá odporúčaný obsah pre tohto používateľa. Techniky založené na modeli (angl. *Model-based Collaborative filtering*) poskytujú odporúčania prostedníctvom oceňovania parametrov štatistických modelov pre používateľské hodnotenia.

4.1.2 Filtrovanie na základe obsahu

Odporúčanie čisto prostedníctvom kolaboratívneho filtrovania využíva iba používateľské hodnotenia. Tieto prístupy berú všetkých používateľov a položky ako atomické jednotky a odporúčania sú vytvárané bez ohľadu na konkrétne špecifiká individuálnych používateľov alebo položiek.

Metódy využívajúce filtrovanie na základe obsahu naopak vytvárajú odporúčania na základe porovnávania modelov reprezentujúcich obsah s modelmi reprezentujúcimi konkrétneho používateľa [10]. Odporúčania v takýchto prístupoch vznikajú na základe prekryvu týchto dvoch modelov.

4.2 Odporúčanie v CQA systémoch

V kontexte CQA systémov je problematika odporúčania a odporúčacích systémov častým objektom výskumu [9].

Jedným z hlavných cieľov CQA systémov je poskytnúť pýtajúcemu sa odpoveď na jeho otázku v čo možno najkratšom čase. Rovnako ako v prípade iných systémov založených na používateľmi vytváranom obsahu, aj v prípade CQA systémov miera nového obsahu – nových otázok a odpovedí – neustále narastá. Napriek tomu je tiež možné pozorovať stúpajúci trend nízkej miery zodpovedanosti otázok [7]. Jedným zo spôsobov, ako je možné riešiť túto situáciu, je práve využitie odporúčacích systémov.

V súčasnosti jedným z trendov najmä v úzko zameraných CQA systémoch ako napr. Stack Overflow, je tiež postupný prechod od modelu jednoduchého odpovedania na položené otázky na model povzbudzujúci k vytváraniu dlhodobo hodnotného obsahu vo forme rozsiahlych komunitne spravovaných odpovedí [3, 6] podnecujúcich diskusiu. V tomto prípade je možné využiť odporúčacie systémy ako prostriedok pre odhalenie a prezentovanie otázok a odpovedí, ktoré by mohli používateľa zaujímať a priniesť mu úžitok aj v prípade, že práve nemá rovnaký problém, ako sa vyskytuje v danej otázke [12].

Výskum v oblasti odporúčania v CQA systémoch sa v súčasnosti zameriava hlavne na oblasti odporúčania, smerovania a získavania otázok. Problémom súčasného výskumu v tejto oblasti je nejednoznačnosť a časté zamieňanie týchto výrazov, prípadne nerozlišovanie medzi odporúčaním a smerovaním otázok [9].

4.2.1 Odporúčanie otázok

Odporúčanie otázok (angl. *Question recommendation*) využíva tzv. *pull* prístup, teda na základe (explicitnej či implicitnej) požiadavky používateľa prezentuje zoznam odporúčaných relevantných otázok (alebo obsahu celkovo). Tento prístup využíva štandardnejší tok medzi použitými modelmi – začína sa modelom používateľa, na ktorý sa odporúčací systém pokúša namapovať model relevantného obsahu. Relevancia otázok pre používateľa môže byť identifikovaná rôznymi prístupmi – či už na základe kolaboratívneho filtrovania (kap. 4.1.1) alebo filtrovania na základe obsahu (kap. 4.1.2).

Forma prezentovania odporúčaných otázok sa tiež môže líšiť. Časté je napríklad zobrazenie otázok, ktoré by používateľa mmohli zaujímať, v detaile konkrétnej otázky, ktorú momentálne používateľ číta. Odporúčanie otázok je však možné využiť aj ako prostriedok pre zvýšenie záujmu a angažovanosti používateľa o CQA systém.

4.2.2 Smerovanie otázok

Na rozdiel od pomerne štandardného odporúčania otázok, v prípade smerovania otázok (angl. *Question routing*) je prístup k odporúčaniu presne opačný, a využíva tzv. *push* prístup. V tomto prípade proces odporúčania začína modelom nezodpovedanej otázky, ktorú sa snaží odporúčací systém nasmerovať k používateľom, ktorí majú najväčší potenciál na túto otázku zodpovedať.

Výskum smerovania nezodpovedaných otázok na konkrétnych používateľov – odpovedajúcich – síce ukazuje, že ide o dôležitý koncept aj z pohľadu používateľského zážitku [13, 14], no prináša so sebou aj problémy. Najvýraznejším z nich je zahltenie expertov, ktorí sú hlavnými terčmi takejto formy odporúčania, nakoľko ich reputácia a expertíza ich predurčuje ako vhodných kandidátov na zodpovedanie veľkého množstva otázok [15].

Pomerne novým prístupom k smerovaniu otázok je namiesto zamerania na konkrétnych používateľov smerovanie otázok na väčšie komunity používateľov [16]. Hlavnou ideou takéhoto smerovania je fakt, že kolektívne poznatky komunity sú vždy väčšie, ako poznatky konkrétneho používateľa, aj experta [17]. Navyše takéto smerovanie zvyšuje pravdepodobnosť rýchlejšieho zodpovedania otázky, ako aj zabraňuje zahlteniu expertov. Hlavným problémom smerovania na komunity je vytváranie kolektívneho modelu reprezentujúceho komunitu, kedy je potrebné brať do úvahy okrem iného fakt, že iba malá časť komunity sú *tvorcovia poznatkov* a nie len ich konzumenti [15].

4.2.3 Získavanie otázok

Tento pojem (angl. *Question retrieval*) v kontexte CQA systémov hovorí o procese vyberania podobných otázok pre rôzne formy dopytov [18] na základe syntaktickej podobnosti otázok. Tento proces je možné využiť na hľadanie odpovedí alebo poznatkov vo veľkých množstvách už zodpovedaných otázok.

Výskum v oblasti získavania otázok v kontexte CQA ststémov sa zameriava hlavne na problém premostenia lexikálnej bariéry medzi obsahovo podobnými otázkami, ktoré sú však formulované použitím iných slov, synonymných výrazov a pod. Na prekonanie tohto problému sa štandardne využíva vytvorenie prekladových modelov medzi otázkami a odpoveďami [19]. Základnou myšlienkou za týmto postupom je predpoklad, že otázky a odpovede sú v podstate *paralelnými textami* a vzťahy medzi nimi môžu byť určené na základe pravdepodobností medzislovných prekladov.

Odlišný prístup k získavaniu otázok v CQA systémoch volia autori v [18]. Argumentujú, že základný predpoklad paralelnosti medzi otázkami a odpoveďami je v praxi nesprávny, pričom problémové sú hlavne odpovede veľmi nízkej kvality. Autori preto navrhujú metódu tématicky-založeného jazykového modelu (angl. *Topic-based Language Model*), ktorá predpokladá, že napriek tomu, že otázky a odpovede sú rozdielne vo viacerých aspektoch, zdieľajú určité spoločne latentné faktory, ktoré predstavujú latentnú tému danej otázky a odpovede. Samotné získavanie otázok je následne postavené práve na modeli, ktorý okrem lexikálnej podobnosti a prekladu berie do úvahy tieto latentné témy.

4.2.4 Problém studeného štartu

Častým problémom odporúčania je problém studeného štartu (angl. *Cold start*), kedy je na dosiahnutie primeranej miery presnosti odporúčania potrebné veľké množstvo informácií, ktoré ale napr. v prípade nových alebo menej aktívnych používateľov nemusia byť k dispozícii.

Tento problém sa vyskytuje najmä v prípade systémov, ktoré obsah odporúčajú na základe podobnosti používateľov medzi sebou. Keďže je často na začiatok potrebné veľké množstvo informácií o daných používateľoch, nie je možné jednoducho odporúčať vhodný obsah pre používateľov, ktorí sú menej aktívni, alebo sú noví. V menšej miere týmto problémom trpia systémy, ktoré namiesto podobnosti používateľov využívajú pre zostavovanie odporúčaní podobnosť samotného obsahu na základe rôznych atribútov.

V kontexte odporúčania otázok v CQA systémoch je tento problém úzko previazaný s problémom dlhého chvosta v používateľskej aktivite (kapitola 3.2.1). Veľmi veľké percento používate-

ľskej základne tvoria používatelia, ktorí sú noví, alebo nemajú žiadnu aktivitu. Pre odporúčanie otázok (kapitola 4.2.1) je tak problém zostaviť profil používateľa, na základe ktorého sa vykonáva mapovanie na model relevantného obsahu. V prípade smerovania otázok (kapitola 4.2.2) je zase problém zostaviť model reprezentujúci expertízu daného používateľa.

4.2.5 Problém filtračnej bubliny

Ďalším problémom, ktorému sa však v oblasti odporúčania obsahu CQA systémov venuje menej pozornosti [9], je rôznorodosť odporúčaného obsahu. Hrozí tak výskyt problému tzv. filtračnej bubliny (angl. *Filter bubble*).

Ak totiž systém používateľovi odporúča obsah len z oblastí používateľovho záujmu, dochádza k problému, kedy je používateľ do značnej miery uzatvorený v rámci jednej oblasti a nemá tak možnosť získavať zaujímavé poznatky z iných oblastí. Používateľ sa tak síce môže stať odborníkom na danú oblasť, no jeho povedomie o širšom kontexte celej problematiky je veľmi obmedzené.

Riešením tohto problému je identifikácia oblastí, ktoré nie sú priamo oblasťami záujmu používateľa, no sú k týmto oblastiam v určitých aspektoch príbuzné. Používateľ má tak možnosť rozšíriť svoj okruh záujmu a vedomosti o širšom kontexte problémovej domény.

4.3 Existujúce metódy pre odporúčanie otázok v CQA

TODO

4.3.1 Črty a algoritmy

Survey

4.3.2 Metodológie overenia

Survey

4.3.3 Problém studeného štartu

- Srba - Utilizing non-QA data to improve questions routing for users with low QA activity in CQA - Cold-Start Expert Finding in Community Question Answering via Graph Regularization

4.3.4 Problém filtračnej bubliny

- Survey - Szpektor - diverzifikacia

4.4 Diskusia

TODO

5 Diverzita a aktuálnosť odporúčania

Diverzitu možno všeobecne definovať ako opak podobnosti. V niektorých prípadoch však nemusí byť odporúčanie podobných položiek tým najlepším riešením pre používateľa [10]. Dôvodom je práve náchylnosť takéhoto odporúčania na vyvolanie problému filtračnej bubliny (viď. kapitola 4.2.5).

Okrem diverzifikácie odporúčaného obsahu má na celkovú úspešnosť vytvárania personalizovaných odporúčaní veľký vplyv aj aktuálnosť (angl. *freshness*, príp. *novelty*) odporúčaného obsahu [20].

5.1 Diverzita v odporúčacích systémoch

Tématická diverzifikácia je metóda napomáhajúca vyváženosti a diverzite personalizovaného odporúčania s cieľom lepšie reflektovať kompletné spektrum používateľových záujmov. Napriek tomu, že môže mať negatívny vplyv na priemernú správnosť odporúčaní, dosahuje táto metóda zvýšenú úroveň používateľskej spokojnosti [21].

Diverzita na základe nízkej vnútornej podobnosti zoznamu
Ziegler a kol. [22] študovali diverzifikáciu v oblasti odporúčaní a navrhli prístup, ktorý vytvára zoznamy odporúčaní lepšie uspokojujúce používateľove záujmy prostredníctvom selekcie takých zoznamov, ktoré majú nízku vnútornú podobnosť.

Odporúčanie bolo navrhnuté s použitím kolaboratívneho filtrovania na základe položiek (kapitola 4.1.1). Vnútorná miera podobnosti položiek zoznamov bola určovaná prostredníctvom metriky založenej na taxonomickej klasifikácií jednotlivých položiek. Samotná diverzifikácia spočívala v sekvenčnom výbere položiek z kandidátnych zoznamov odporúčaní tak, aby bola minimalizovaná vnútorná tématická podobnosť výsledného zoznamu odporúčaní. Autori v online experimente demonštrovali, že reálni používatelia preferujú diverznejšie výsledky.

Diverzita na základe dôvodu
Tradičný spôsob zavedenia diverzity do odporúčania je diverzifikácia na základe atribútov odporúčaného obsahu, teda zoskupenie výsledkov do skupín zdieľajúcich viaceré atribúty (ako napr. žáner hudby) a následný výber iba limitovaného množstva výsledkov z každej zo skupín. Autori v [23] prezentujú *diverzifikáciu na základe dôvodu*. Táto metóda využíva pre diverzifikáciu výsledkov dôvod, prečo bola konkrétna položka odporučená (napr. *tento album bol odporu-*

čený, pretože ste počúvali inú skladbu tohto autora). V článku autori nekonkretizujú spôsob určenia dôvodov pre odporúčanie, len formálne definujú metriku pre výpočet diverznosti medzi odporúčanými položkami ako priemer kosínovej vzdialenosti medzi vektormi reprezentujúcimi tieto dôvody. Pre aplikovanie diverzifikácie v odporúčaní *top-K* položiek využívajú nasledovný algoritmus:

> Pre daného používateľa u a prah θ pre agregované skóre množiny odporúčaných položiek σ, nájdi množinu $S \subseteq \sigma$ takú, že $|S| = k, score(S) >= \theta$ a priemerná kosínová vzdialenosť diverzity položiek je maximalizovaná.

Diverzifikáciu na základe dôvodu autori porovnávajú so štandardným prístupom diverzifikácie na základe atribútov odporúčaného obsahu. Autori experimentálne ukázali, že takáto forma diverzifikácie je prinajmenšom rovnako účinná, ako diverzifikácia na základe atribútov položiek, pričom z pohľadu výkonu ju výrazne presahuje.

Diverzita na báze proporcionality
Inú perspektívu volí metóda diverzity na báze proporcionality. Zoznam odporúčaní je možné považovať za najlepšie diverzifikovaný vzhľadom na relevanciu odporúčaní v takom prípade, keď počet výsledkov z určitej témy je úmerný popularite danej témy. Dang a kol. vo svojej práci [24] ponúkajú koncept optimalizácie proporčnosti pre diverzifikáciu výsledkov vyhľadávania.

Napriek tomu, že sa práca [24] nezaoberá priamo diverzitou v odporúčacích systémov, sú paralely s touto oblasťou výrazné. Motiváciou pre tákyto spôsob diverzifikácie je metóda obsadzovania kresiel v parlamente. Ich metóda postupne pre každú pozíciu v zozname výsledkov určuje tému, ktorá najlepšie zachováva celkovú proporčnosť. Následne na túto pozíciu z danej témy vyberie najlepší dokument.

5.2 Aktuálnosť v odporúčacích systémoch

Aktuálnosť v kontexte odporúčacích systémov môže predstavovať dva rôzne aspekty. Jedným z nich je *novosť* (angl. *novelty*), teda pomerne priamočiara vlastnosť určujúca, či odporúčaný obsah už bol používateľovi prezentovaný, alebo nie. Jednoduchým spôsobom, ako zabezpečiť, aby používateľovi neboli stále dookola odporúčané tie isté položky, akokoľvek relevantné by pre neho mohli byť, je odfiltrovanie položiek, ktoré už používateľovi boli odporúčané v minulosti, a s ktorými už interagoval [10].

Druhým aspektom aktuálnosti je *čerstvosť* (angl. *freshness*). V tomto prípade ide o časovú aktuálnosť odporúčaného obsahu. Naivný prístup k aktuálnosti je odporúčanie iba obsahu z určitého obmedzeného časového úseku z blízkej minulosti, no takýto prístup nemusí vždy

dosahovať najlepšie výsledky používateľskej spokojnosti [25].

Pri odporúčaní, ktoré berie do úvahy aktuálnosť odporúčaného obsahu je dôležitým aspektom detekcia časovo citlivej témy. Takýto systém by mal presadzovať aktuálny obsah iba v prípade, kedy je to vhodné. Na druhej strane, ak je v prípade aktívne sa vyvýjajúcej témy odporúčaný neaktuálny obsah, môže to výrazne degradovať úspešnosť odporúčania [26]. Ďalším faktorom pri posudzovaní aktuálnosti v odporúčaní je časová škála aktuálnosti pre danú tému. V prípade niektorých tém alebo oblastí je možné považovať za aktuálne položky z posledného roka, no v iných prípadoch môžu byť aj niekoľko týždňov staré položky považované za vysoko neaktuálne.

Ďalším problémom vzhľadom na aktuálnosť v odporúčaní je tiež fakt, že pre novo vzniknutý obsah môže byť problémovejšie zostaviť model, ktorý by ho reprezentoval, nakoľko môže byť o tomto obsahu známych zatiaľ iba málo informácií [27]. Tento problém studeného štartu sa môže prejavovať rovnako v prípade modelovania obsahu, ako je tomu napríklad v prípade vytvárania modelov reprezentujúcich nových používateľov.

Riešením v takomto prípade môže byť napríklad vytváranie modelu obsahu na základe čŕt, ktoré nie sú ovplyvnené časom (napr. nadpis, text alebo autor obsahu, na rozdiel od počtu hlasov alebo dátumu uverejnenia), alebo tiež upravenie hodnôt týchto čŕt vzhľadom na relatívnu aktuálnosť obsahu.

5.3 Diverzita a aktuálnosť v kontexte CQA systémov

Kým diverzita a aktuálnosť sú v oblasti odporúčania a celkovo vo vyhľadávaní informácií pomerne často analyzovanými aspektmi, v kontexte CQA systémov sa týmto hľadiskám doteraz venovala iba okrajová pozornosť [9].

Liu a kol. vo svojej práci [20] skúmajú aspekt aktuálnosti odporúčania v CQA systémoch prostredníctvom relatívne neštandardného návrhu CQA systému určeného pre odpovedanie v reálnom čase na *hyper-lokálne* a časovo senzitívne otázky. Za týmto účelom využívajú prístupy predchádzajúcich prác a kombinujú aspekty relevancie, lokality a aktuálnosti v *real-time* CQA systéme.

Komplexnejší pohľad na aktuálnosť a diverzitu priamo v kontexte štandardných CQA systémov ponúka Szpektor a kol [25]. Autori experimentovali so zavedením diverzity a aktuálnosti do procesu vytvárania odporúčaní pre používateľov CQA systému Yahoo! Answers, pričom sa na rozdiel od väčšiny prác v tejto oblasti nezameriavali len na skupinu expertných používateľov.

Pre odporúčanie využili profil otázok založený na kombinácii LDA, lexikálneho a kategorického

modelu a profil používateľa odvodený od profilu otázok, s ktorými interagoval. Párovanie otázok a používateľov bolo vykonané prostredníctvom jednoduchého skalárneho súčinu vektorov reprezentujúcich profily používateľov a otázok.

Samotná diverzifikácia odporúčaní bola vykonávaná prostredníctvom tématického výberu vzoriek (angl. *thematic sampling*), kedy je vygenerovaných viacero samostatných zoznamov odporúčaných otázok z viacerých tém, ktoré sú následne zmiešané dokopy proporcionálne k pravdepodobnostnému skóre jednotlivých tématických zoznamov.

Prínos aktuálnosti do odporúčania v CQA systémoch bol skúmaný na základe odporúčania iba aktuálneho obsahu – konkrétne iba nezodpovedaných otázok za posledné štyri hodiny.

Dopad diverzifkácie a aktuálnosti na úspešnosť odporúčania bol testovaný v rámci online experimentu. Používatelia boli náhodne rozdelení do štyroch segmentov:

1. **Kontrolná vzorka** – Týmto používateľom neboli ponúknuté žiadne odporúčania.
2. **Odporúčanie na základe relevancie** – Používateľom boli ponúknuté odporúčania iba na základe relevancie daných otázok, bez ohľadu na aktuálnosť alebo diverzitu.
3. **Odporúčanie s ohľadom na aktuálnosť** – Používateľom boli odporúčané relevantné otázky, pričom 50% z nich pochádzalo z posledných štyroch hodín a 20% bolo vybraných prostredníctvom tématického výberu vzoriek.
4. **Diverzifikované odporúčanie** – Používateľom boli odporúčané relevantné otázky, pričom 50% z nich bolo vybraných na základe tématického výberu vzoriek ako prostriedku diverzifikácie, a 20% pochádzalo z posledných štyroch hodín.

Výsledky online experimentu potvrdili intuitívnu myšlienku, že iba samotná relevancia nie je dostatočná na úspešne odporúčanie otázok v CQA systéme. Práve naopak, vo vykonanom experimente dokonca samotné odporúčanie len na základe relevancie dosiahlo nižšie hodnoty zodpovedania otázok, ako kontrolná vzorka bez akýchkoľvek odporúčaní.

Presadzovanie aktuálnych otázok dosiahlo zvýšenie miery zodpovedania otázok o 4%, avšak najlepšie výsledky boli dosiahnuté prostredníctvom diverzifikácie odporúčaní aj za cenu zníženia aktuálnosti, pričom miera zodpovedania sa zvýšila o 17%.

5.4 Diskusia

Na základe tejto analýzy môžeme usúdiť, že napriek tomu, že problematika diverzifikácie a aktuálnosti odporúčania v kontexte CQA systémov v súčasnosti stále ostáva do veľkej miery

nepreskúmaná, je očividné, že uvažovanie týchto aspektov v tomto kontexte má veľký vplyv na úspešnosť odporúčania, pričom informačné bulletiny sa javia ako prirodzená, požadovaná, no napriek tomu málo využívaná forma prinášania odporúčaného a potenciálne zaujímavého obsahu používateľom.

6 Návrh riešenia personalizovaného informačného bulletinu v CQA systéme

Cieľom našej práce je navrhnúť, zrealizovať a overiť metódu zostavovania personalizovaných informačných bulletinov v CQA systémoch so zameraním sa na podporu diverzity a aktuálnosti obsahu odporúčaného v informačnom bulletine.

Naše riešenie je navrhnuté pre použitie vrámci platformy Stack Exchange, ktorá patrí medzi najpopulárnejšie CQA systémy v súčasnosti a tvorí ju viac ako 160 komunít zameraných na rôzne oblasti.

Prehľad navrhovanej metódy
Na začiatok uvádzame stručný súhrn navrhnutej metódy zostavovania personalizovaných informačných bulletinov. Celkový pohľad na metódu ilustruje obrázok 6, Jednotlivé časti metódy sú podrobne opísané v nasledujúcich kapitolách.

- Pre zostavenie personalizovaných informačných bulletinov využijeme odporúčaciu metódu filtrovania na základe obsahu.
- Diverzifikáciu odporúčaného obsahu budeme zabezpečovať na úrovni tématického prefilteru pred samotným procesom odporúčania – pre každú nezávislú tému sa bude vytvárať samostatný zoznam odporúčaní.
- Výsledný zoznam odporúčaného obsahu vznikne spojením jednotlivých zoznamov z kroku diverzifikácie.
- Pri odporúčaní sa okrem diverznosti a relevancie obsahu bude uvažovať aj jeho aktuálnosť.

TODO - sem pojde flow diagram znazorujuci cely proces zostavovania odporucani

6.1 Návrh metódy personalizovaného odporúčania

Hypotéza
Použitím personalizovaného odporúčania otázok v informačnom bulletine CQA systému zvýšime relevanciu obsahu informačného bulletinu, čo sa prejaví zvýšenou mierou jeho používania medzi používateľmi CQA systému.

Pri vytváraní personalizovaného informačného bulletinu sa budeme zameriavať na odporúča-

nie relevantných otázok jednotlivým používateľom CQA systému prostredníctvom aplikovania metódy filtrovania na základe obsahu (kapitola 4.1.2). Pre účely filtrovania na základe obsahu je potrebné definovať a zostaviť modely reprezentujúce jednak otázky, a tiež používateľov CQA systému.

6.1.1 Model otázok

Model otázky sa bude skladať z troch nezávislých modelov, ktoré sa na samotnú otázku pozerajú z rôznych perspektív.

1. Kategorický model otázok
Tento model reprezentuje otázku na najvyššej úrovni ako prislúchajúcu do určitých kategórií. Kategórie otázok sú vrámci platformy Stack Exchange reprezentované ako značky (angl.*tags*). Každá otázka môže obsahovať viacero značiek.

Samotný kategorický model otázky bude reprezentovaný ako vektor v n-rozmernom priestore, kde každý rozmer k predstavuje príslušnosť otázky k danej značke. Nakoľko značky netvoria hierarchickú štruktúru, bude vektor v jednotlivých rozmeroch nadobúdať iba hodnoty 0 alebo 1.

2. Tématický model otázok
Tématický model využíva metódu latentnej Dirichletovej alokácie (angl. *Latent Dirichlet Allocation - LDA*) [28] na určenie latentnej témy, ktorej sa daná otázka venuje.

LDA vektor tohto modelu bude reprezentovať distribúciu otázky vrámci jednotlivých latentných tém. Samotné LDA témy sa budú odvodzovať z nadpisu a textu otázky. Pre optimalizáciu modelu a zanedbanie tém s veľmi nízkou distribúciou sa do úvahy budú brať len latentné témy tvoriace 75% z celkovej distribúcie, teda tretí kvartil. Jednotlivé hodnoty tém budú následne normalizované, aby tvorili 100%.

Nastavenie LDA
Pre určenie vhodného počtu LDA tém (n) využijeme metódu hierarchických Dirichletových procesov [29]. Pre trénovanie LDA modelu bude využitý online variačný Bayesov algoritmus. Parametre modelu budú nastavené nasledovne:

$$\alpha = \frac{1}{n}; \kappa = 0.7; \tau_0 = 10; \eta = \frac{1}{n}$$

Nakoľko jednotlivé komunity vrámci platformy Stack Exchange sú pomerne úzko zamerané, predpokladáme, že bude dostačujúce natrénovať LDA model na vzorke archívnych dát a nebude potrebné postupné dotrénovanie modelu. Napriek tomu

sme zvolili online variačný Bayesov algoritmus, keďže je pri veľkom množstve dát efektívnejší ako dávková varianta tohto algoritmu.

3. Lexikálny model otázok Lexikálny model otázky využíva TF-IDF vektor reprezentujúci zastúpenie jednotlivých výrazov v texte otázky. Rovnako ako LDA vektor bude zostavený z nadpisu a textu samotnej otázky. Pred výpočtom bude text lematizovaný a budú z neho odstránené stop slová.

6.1.2 Model používateľov

Pre každého používateľa budeme uvažovať dva nezávislé modely – jeden bude modelovať záujem používateľa o určité témy a otázky, druhý bude modelovať jeho expertízu v určitej oblasti. Jeden bude použitý pri odporúčaní otázok, ktoré by používateľa mohli zaujímať, druhý pri odporúčaní otázok, ktoré by mohol vedieť zodpovedať. Oba modely budú z pohľadu svojej štruktúry presne rovnaké.

Model používateľa bude zostavený na základe jeho aktivity a bude sa analogicky k modelu otázok skladať z troch vektorov:

1. Prvý vektor bude reprezentovať aktivitu používateľa naprieč značkami – každý rozmer bude predstavovať jednu značku, v ktorej má používateľ aktivitu. Hodnoty v jednotlivých rozmeroch budú predstavovať podiel aktivity v danej značke voči celkovému množstvu aktivity používateľa.
2. Druhý vektor bude analogicky k prvému reprezentovať aktivitu používateľa naprieč LDA témami, v ktorých má používateľ aktivitu.
3. Tretí vektor bude reprezentovať zastúpenie jednotlivých výrazov v textoch otázok, v ktorých má používateľ aktivitu.

Záujmový model
Model predstavujúci záujem používateľa bude zostavený zo všetkých otázok, ktoré používateľ položil, alebo ktoré označil za obľúbené. Každá takáto otázka bude do modelu prispievať rovnakou váhou.

Expertízny model
Model modelujúci expertízu používateľa sa bude skladať z otázok, na ktoré používateľ odpovedal, pričom ich dopad na model expertízy bude závislý od skóre jeho odpovede. Kladné skóre bude prisipevať do modelu pozitívne – bude to teda signalizovať fakt, že používateľ danej téme rozumie. Odpoveď so záporným skóre bude naopak signalizovať, že používateľ danej téme nero-

zumie, čo bude reflektované aj v jeho expertíznom modeli. Odpovede označené za akceptované budú do modelu prispievať s koeficientom $k = 1.5$.

Komentáre
Okrem pokladania otázok, odpovedania alebo označenia za obľúbené budú uvažované aj používateľove komentáre. Keďže však zo samotného faktu že používateľ niečo okomentoval nemožno jednoznačne určiť, či táto aktivita predstavuje jeho záujem alebo expertízu, budú otázky, ktoré používateľ okomentoval prispievať do oboch modelov – záujmového aj expertízneho, avšak s koeficientom $k = \frac{1}{3}$.

6.1.3 Výber odporúčaného obsahu

Pre účely zostavovania zoznamu odporúčaných otázok použijeme prístup analogický štandardným nástrojom pre vyhľadávanie informácií (angl. *Information Retrieval Engines*). V našom prípade budú dokumentmi samotné otázky a dopytom bude model používateľa. Pre ohodnocovanie podobnosti modelov bude použitý skalárny súčin vektorov reprezentujúcich modely otázky a používateľa.

6.1.4 Riešenie problému studeného štartu

Pre eliminovanie problému studeného štartu (kapitola 4.2.4) z pohľadu prvotného odporúčania otázok využijeme offline natrénovanie našich metód na archívnych dátach platformy Stack Exchange.

V prípade nových používateľov, ktorí v systéme nemajú žiadnu, alebo iba nedostatočnú aktivitu, budeme zo začiatku vytvárať iba generický informačný bulletin podobný tomu, ktorý je používateľom k dispozícii aj v súčasnosti (kapitola 3.3.1).

6.1.5 Aktuálnosť / Freshness

TODO - exponencialny decay factor, zavisly od priemernej frekvencie userovej aktivity. Plus berieme len content za max 2x frekvencie odosielania newsletteru.

6.2 Návrh metód diverzifikácie odporúčaní

Hypotéza
Výberom odporúčaných otázok zo širšieho okruhu záujmu používateľa a zohľadnením ich aktuálnosti predídeme výskytu problému filtračnej bubliny, čím dosiahneme vyššiu mieru záujmu používateľa a jeho aktivity.

Diverzifikácia obsahu personalizovaného informačného bulletinu sa bude vykonávať ešte pred samotným zostavovaním zoznamu odporúčaní formou pre-filteru. Diverzifikácia bude spočívať vo výbere *značiek* a *tém*, pričom následne v kroku zostavovania zoznamov odporúčaní sa bude pre každý zoznam uvažovať len obsah prislúchajúci do danej značky alebo témy. Okrem výberu značiek a tém bude diverzifikácia aplikovaná aj pri následnom výbere položiek z jednotlivých zoznamov do výsledného zoznamu odporúčaní.

Tento prístup k diverzifikácii sme zvolili okrem jeho prirodzenosti aj z dôvodu jeho veľmi dobrej škálovateľnosti, nakoľko alternatívny prístup postavený na tvorbe odporúčaní nad všetkým obsahom daného CQA systému a až následnej diverzifikácií by v praxi nebol realizovateľný.

Cieľom našej práce je vyhodnotiť dopad diverzifikácie odporúčaní na personalizované informačné bulletiny CQA systémov. Za týmto účelom sme navrhli dve metódy diverzifikácie odporúčaní: metódu proporčnej diverzifikácie a metódu tématického vzorkovania.

6.2.1 Metóda proporčnej diverzifikácie

Zostavenie výsledného zoznamu n odporúčaní s použitím metódy proporčnej diverzifikácie je navrhnuté nasledovne:

1. Pre každého používateľa vyberieme z jeho záujmového alebo expertízneho modelu k najvyššie hodnotených značiek z kategorického vektoru a k najvyššie hodnotených tém z tématického vektoru, pričom $2 \times k = \left\lceil \frac{n}{2} \right\rceil$.
2. Prostredníctvom vyššie opísanej metódy (kapitola 6.1.3) sa pre každú takúto značku a tému následne zostaví zoznam n odporúčaní, pričom sa budú uvažovať len otázky prislúchajúce tejto téme alebo značke.
3. Následne sa z každého zoznamu vyberie vrchných n_k odporúčaných položiek, pričom $n_k = \frac{n}{2k}$

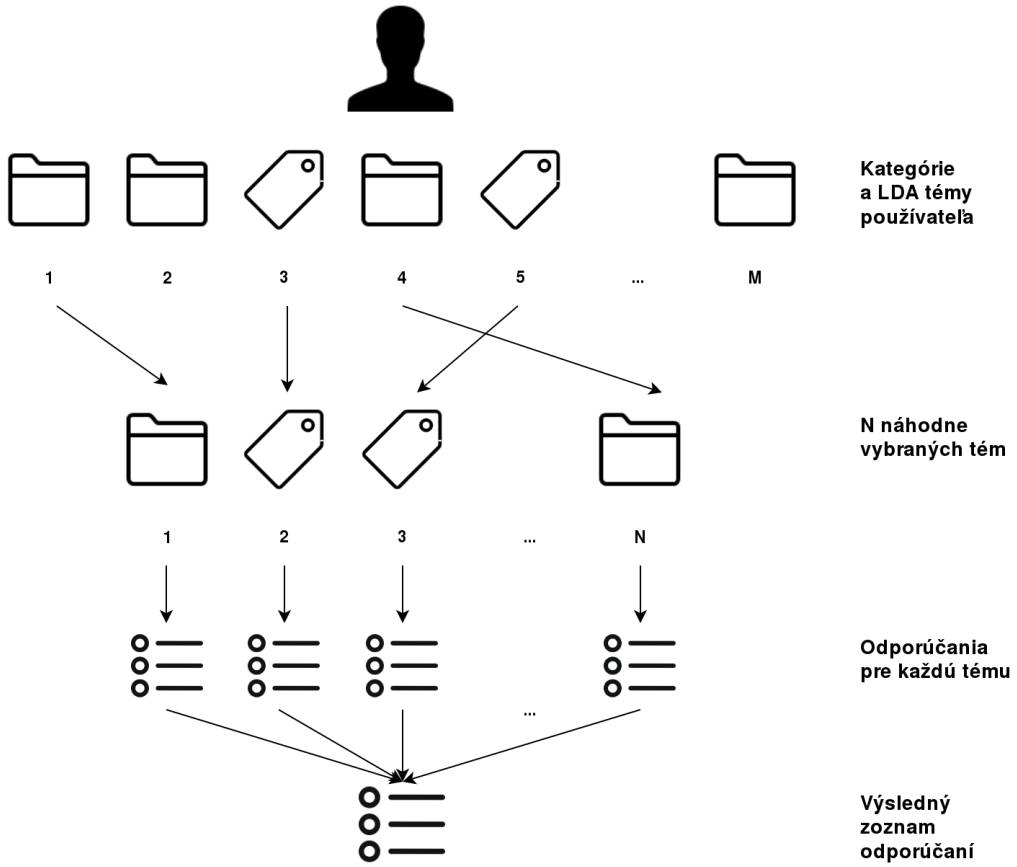
Metóda proporčnej diverzifikácie tak v podstate používateľovi vyberie najrelevantnejšie položky z jeho najrelevantnejších tém a značiek. T'uto jednoduchú metódu diverzifikácie budeme

následne porovnávať s metódou tématického vzorkovania.

6.2.2 Metóda tématického vzorkovania

Pri návrhu tejto metódy sme sa inšpirovali metódou rovnomennou metódou z [25], ktorú sme prispôsobili podmienkam špecifickým pre našu prácu. Proces tématického vzorkovania ilustruje schéma na obrázku 6.1.

1. Pre každého používateľa vyberieme z jeho modelu k_1 značiek a k_2 tém. Pri výbere sa obmedzíme na značky a témy, ktoré spadajú do druhého kvartilu – mediánu. Potlačíme tak značky a témy, ktoré majú pre používateľa nízku relevanciu.
   $k_1 + k_2 = \left\lceil \frac{n}{2} \right\rceil$; konkrétne hodnoty pre k_1 a k_2 nešpecifikujeme, n je počet odporúčaní vo výslednom zozname.
2. Prostredníctvom vyššie opísanej metódy (kapitola 6.1.3) sa pre každú takúto značku a tému následne zostaví zoznam n odporúčaní, pričom sa budú uvažovať len otázky prislúchajúce tejto téme alebo značke.
3. Následne budeme náhodne vyberať vzorky z jednotlivých zoznamov do výsledného zoznamu odporúčaní, pričom pravdepodobnosť výberu otázky z konkrétneho zoznamu $P_s(l_i)$ bude proporčná k relevancii tejto témy pre daného používateľa.
4. Konkrétne otázky z jednotlivých zoznamov sa budú vyberať z $m = P_s(l_i)$ najrelevantnejších otázok daného zoznamu v náhodnom poradí.



Obr. 6.1: *Schéma metódy diverzifikácie odporúčaní tématickým vzorkovaním.*

6.3 Metriky hodnotenia výsledkov

Pre overovanie výsledkov experimentov budeme používať tieto metriky:

TODO pridat vzorce pre jednotlive metriky

Precision@N

Presnosť (angl. *Precision*), alebo tiež *pozitívna predikčná hodnota* je metrika reprezentujúca pomer relevantných dokumentov z celkového zoznamu. Štandardne sa presnosť počíta ako pomer z celého zoznamu dokumentov, no v oblasti odporúčania a vyhľadávania informácií je často vhodnejšou odvodená metrika *Precision@N*, ktorá určuje, aká časť z prvých N dokumentov v zozname je relevantná.

nDCG

Normalized Discounted Cumulative Gain je metrika kvality ohodnocovania často používaná na meranie efektívnosti odporúčania. DCG meria užitočnosť dokumentov na základe ich pozície

vo výslednom zozname. Užitočnosť dokumentov sa akumuluje od konca zoznamu, pričom najvyššiu užitočnosť majú dokumenty na začiatku zoznamu [30].

CTR
Miera preklikov (angl. *Click-through Rate*) je metrika často využívaná v spojitosti s informačnými bulletinmi. Táto metrika vyjadruje počet úspešných kliknutí na odkaz v informačnom bulletine.

Okrem CTR plánujeme v súvislosti s interakciou používateľa s informačným bulletinom merať aj počet impresií, teda zobrazení informačného bulletinu, ako aj počet konverzií, teda podiel prípadov, kedy kliknutie na niektorú z otázok v informačnom bulletine viedlo k aktivite používateľa na tejto otázke – či už označenie za obľúbenú, odpovedanie alebo pridanie komentáru. Ďalej tiež plánujeme merať počet odhlásení z informačného bulletinu.

6.4 Návrh overenia metód

Nami navrhnuté metódy personalizovaného odporúčania a diverzifikácie odporúčaného obsahu budeme overovať prostredníctvom online nekontrolovaného experimentu spoločne s kolegom Matúšom Salátom [31] na používateľoch z komunity *Stack Overflow* platformy Stack Exchange.

Tento online experiment bude mať formu pravidelne rozposielaného informačného bulletinu, na ktorého odoberanie sa budú môcť prihlásiť všetci používatelia z tejto komunity.

Účinnosť zvolených metód diverzifikácie odporúčaní budeme vyhodnocovať prostredníctvom A/B testovania, pričom používateľov rozdelíme na tri skupiny:

1. Kontrolná skupina – odporúčania nebudú diverzifikované
2. Skupina A – využitie metódy tematického vzorkovania
3. Skupina B – využitie metódy proporčnej diverzifikácie

Dôležitým predpokladom pre úspešnosť online experimentu bude získať dostatočnú reprezentatívnu vzorku používateľov ochotných byť súčasťou experimentu. V prípade, že by sa nám nepodarilo osloviť dostatočný počet používateľov, plánujeme vykonať kontrolovaný offline experiment na vzorke archívnych dát.

Literatúra

[1] Silverpop Systems. Email marketing metrics benchmark study. *White paper*, Silverpop Systems, Inc., 2012.

[2] Qiaoling Liu, Eugene Agichtein, Gideon Dror, Yoelle Maarek, and Idan Szpektor. When web search fails, searchers become askers. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12*. ACM Press, 2012.

[3] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Discovering value from community activity on focused question answering sites. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*. ACM Press, 2012.

[4] Alton Y. K. Chua and Snehasish Banerjee. Where to ask and how to ask? the case of community question answering sites. In *2014 Science and Information Conference*. IEEE, aug 2014.

[5] Jing Li, Zhenchang Xing, Deheng Ye, and Xuejiao Zhao. From discussion to wisdom. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing - SAC '16*. ACM Press, 2016.

[6] Guo Li, Haiyi Zhu, Tun Lu, Xianghua Ding, and Ning Gu. Is it good to be like wikipedia? In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*. ACM Press, 2015.

[7] Ivan Srba and Maria Bielikova. Why is stack overflow failing? preserving sustainability in community question answering. *IEEE Software*, 33(4):80–89, jul 2016.

[8] Luca Ponzanelli, Andrea Mocci, Alberto Bacchelli, Michele Lanza, and David Fullerton. Improving low quality stack overflow post detection. In *2014 IEEE International Conference on Software Maintenance and Evolution*. IEEE, sep 2014.

[9] Ivan Srba and Maria Bielikova. A comprehensive survey and classification of approaches for community question answering. *ACM Transactions on the Web*, 10(3):1–63, August 2016.

[10] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors. *Recommender Systems Handbook*. Springer US, 2011.

[11] M. D. Buhmann, Prem Melville, Vikas Sindhwani, Novi Quadrianto, Wray L. Buntine, Luís Torgo, Xinhua Zhang, Peter Stone, Jan Struyf, Hendrik Blockeel, Kurt Driessens, Risto Miikkulainen, Eric Wiewiora, Jan Peters, Russ Tedrake, Nicholas Roy, Jun Morimoto, Peter A. Flach, and Johannes Fürnkranz. Recommender systems. In *Encyclopedia of Machine Learning*, pages 829–838. Springer US, 2011.

[12] Hapnes Toba, Zhao-Yan Ming, Mirna Adriani, and Tat-Seng Chua. Discovering high quality answers in community question answering archives using a hierarchy of classifiers. *Information Sciences*, 261:101–115, mar 2014.

[13] Wei Li, Charles Zhang, and Songlin Hu. G-finder. In *Proceedings of the ACM international conference on Object oriented programming systems languages and applications - OOPSLA '10*. ACM Press, 2010.

[14] Baichuan Li, Irwin King, and Michael R. Lyu. Question routing in community question answering. In *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*. ACM Press, 2011.

[15] Aditya Pal. Metrics and algorithms for routing questions to user communities. *ACM Transactions on Information Systems*, 33(3):1–29, mar 2015.

[16] Duen-Ren Liu, Yu-Hsuan Chen, and Chun-Kai Huang. QA document recommendations for communities of question–answering websites. *Knowledge-Based Systems*, 57:146–160, feb 2014.

[17] Aditya Pal, Fei Wang, Michelle X. Zhou, Jeffrey Nichols, and Barton A. Smith. Question routing to user communities. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13*. ACM Press, 2013.

[18] Kai Zhang, Wei Wu, Haocheng Wu, Zhoujun Li, and Ming Zhou. Question retrieval with high quality answers in community question answering. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM '14*. ACM Press, 2014.

[19] Xin Cao, Gao Cong, Bin Cui, and Christian S. Jensen. A generalized framework of exploring category information for question retrieval in community question answer archives. In *Proceedings of the 19th international conference on World wide web - WWW '10*. ACM Press, 2010.

[20] Qiaoling Liu, Tomasz Jurczyk, Jinho Choi, and Eugene Agichtein. Real-time community question answering. In *Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI '15*. ACM Press, 2015.

[21] Fuguo Zhang. Improving recommendation lists through neighbor diversification. In *2009 IEEE International Conference on Intelligent Computing and Intelligent Systems*. IEEE, nov 2009.

[22] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web - WWW '05*. ACM Press, 2005.

[23] Cong Yu, Laks V. S. Lakshmanan, and Sihem Amer-Yahia. Recommendation diversification using explanations. In *2009 IEEE 25th International Conference on Data Engineering*. IEEE, mar 2009.

[24] Van Dang and W. Bruce Croft. Diversity by proportionality. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12*. ACM Press, 2012.

[25] Idan Szpektor, Yoelle Maarek, and Dan Pelleg. When relevance is not enough. In *Proceedings of the 22nd international conference on World Wide Web - WWW '13*. ACM Press, 2013.

[26] Anlei Dong, Yi Chang, Zhaohui Zheng, Gilad Mishne, Jing Bai, Ruiqiang Zhang, Karolina Buchner, Ciya Liao, and Fernando Diaz. Towards recency ranking in web search. In *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10*. ACM Press, 2010.

[27] Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, and Hongyuan Zha. Time is of the essence. In *Proceedings of the 19th international conference on World wide web - WWW '10*. ACM Press, 2010.

[28] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[29] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, dec 2006.

[30] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, oct 2002.

[31] Matúš Salát. Todo - doplnit nazov prace. Master's thesis, Fakulta informatiky a informačných technológií, STU Bratislava, 2018.

A Plán práce na diplomovom projekte

A.1 Plán práce na diplomovom projekte I

Prácu na diplomovom projekte I sme navrhli a naplánovali nasledovne:

Tabuľka A.1: *Plán práce na diplomovom projekte I*

1-5. týždeň semestra	Analýza problematiky odporúčania v kontexte CQA systémov a súčasného výskumu v tejto oblasti.
6-7. týždeň semestra	Analýza dostupných dát na platforme Stack Exchange a možností verejného API platformy.
8-9. týždeň semestra	Vytvorenie predbežného návrh metódy riešenia problému a návrh metód overenia výsledkov.
10-12. týždeň semestra	Spísanie správy diplomového projektu I v rátane analýzy a predbežného návrhu riešenia a overenia.

Zhodnotenie

Navrhnutý plán práce v rámci diplomového projektu I sa nám do veľkej miery podarilo dodržať. Časový sklz sa prejavil až vo fáze návrhu metódy riešenia problému, čím sa posunulo spísanie správy až do 11. týždňa semestra.

A.2 Plán práce na diplomovom projekte II

Tabuľka A.2: *Plán práce na diplomovom projekte II*

1-2. týždeň semestra	– Príprava databázy a aktualizačného modulu. – Vytvorenie modelov pre získavanie spätnej väzby od používateľov. – Prvotná dátová analýza.
3-4. týždeň semestra	– Príprava rozhrania pre odoberanie informačných bulletinov. – Generovanie informačných bulletinov na základe prvotného modelu. – Nasadenie systému a otestovanie v reálnych podmienkach.
5-8. týždeň semestra	– Získavanie používateľov informačného bulletinu. – Spracovanie dát, vytvorenie reálnych modelov používateľov a otázok, generovanie odporúčaní. – Príprava metód diverzifikácie odporúčaní. – Písanie diplomovej práce.
9-12. týždeň semestra	– Overenie a vyhodnocovanie informačných bulletinov. – Nasadenie a porovnanie metód diverzifikácie. – Počiatočné overenie výsledkov. – V prípade neúspechu online experimentu, plánovanie offline expermientov.

A.3 Plán práce na diplomovom projekte III

Tabuľka A.3: *Plán práce na diplomovom projekte III*

1. mesiac	Pokračovanie v experimentoch, nasadzovanie systému na väčšom množstve dát.
2. mesiac	Analýza výsledkov experimentov, vyhodnotenie úspešnosti projektu.
3. mesiac	Dokončenie diplomovej práce.