

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-5208-72397

Bc. Martin Šrank

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

Diplomová práca

Vedúci práce: Ing. Ivan Srba, PhD.

Máj 2018

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

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 2018

Anotácia

Slovenská technická univerzita v Bratislave

FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLOGIÍ

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 2018

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.

TODO – Doplniť a trochu rozpisat (vid doc.)

Annotation

Slovak University of Technology Bratislava

FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES

Degree Course: Information Systems

Author: Martin Šrank

Master thesis: Improving Diversity and Freshness of Newsletters in Community Question Answering Systems

Supervisor: Dr. Ivan Srba

2018, May

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.

TODO – Doplňit a trochu rozpisat (vid doc.)

POĎAKOVANIE

Chcel by som sa touto cestou poďakovať vedúcemu mojej práce, Ivanovi Srbovi, za jeho ochotu a odbornú pomoc pri vypracovaní tejto práce. Tiež by som sa chcel za spoluprácu poďakovať môjmu kolegovi Matúšovi Salátovi, s ktorým sme spoločne pracovali na projekte StackLetter, ako aj všetkým, ktorí sa zúčastnili nášho online nekontrolovaného experimentu.

ČESTNÉ VYHLÁSENIE

Čestne vyhlasujem, že záverečnú prácu som vypracoval samostatne s použitím uvedenej literatúry a na základe svojich vedomostí a znalostí.

V Bratislave, 2. 5. 2018

.....
Bc. Martin Šrank

Obsah

1	Úvod	1
2	Informačné bulletiny	2
2.1	Systémy využívajúce informačné bulletiny	2
2.2	Problémy informačných bulletinov	3
2.3	Diskusia	4
3	CQA systémy a odporúčanie	5
3.1	Druhy CQA systémov	5
3.2	Problémy CQA systémov	6
3.2.1	Problém dlhého chvosta v aktivite používateľov	6
3.2.2	Variabilita v kvalite obsahu	6
3.3	Odporúčacie systémy	7
3.3.1	Kolaboratívne filtrovanie	7
3.3.2	Filtrovanie na základe obsahu	8
3.4	Odporúčanie v CQA systémoch	8
3.4.1	Odporúčanie otázok	9
3.4.2	Smerovanie otázok	9
3.4.3	Získavanie otázok	10
3.4.4	Problém studeného štartu	10
3.4.5	Problém filtračnej bubliny	11
3.5	Informačné bulletiny v CQA systémoch	11
3.5.1	Informačné bulletiny v sieti Stack Exchange	12
3.5.2	Informačný bulletin portálu Quora	14
3.5.3	CQA systémy bez informačných bulletinov	15
4	Diverzita a aktuálnosť odporúčania	16
4.1	Diverzita v odporúčacích systémoch	16
4.2	Aktuálnosť v odporúčacích systémoch	17
4.3	Diverzita a aktuálnosť v kontexte CQA systémov	18
4.4	Diskusia	19
5	Návrh riešenia personalizovaného informačného bulletinu v CQA systéme	21
5.1	Návrh metódy personalizovaného odporúčania	22
5.1.1	Profil otázok	22
5.1.2	Profil používateľov	24

5.1.3	Výber odporúčaného obsahu	26
5.1.4	Riešenie problému studeného štartu	26
5.1.5	Aktuálnosť odporúčaní	26
5.2	Návrh metódy diverzifikácie odporúčaní	27
5.3	Zhrnutie	29
6	StackLetter – personalizovaný informačný bulletin pre Stack Exchange	30
6.1	Prehľad modulov	31
6.2	Použité technológie a služby	32
6.3	Realizácia personalizovaného odporúčania	33
6.3.1	Zostavenie LDA a TF-IDF modelov	33
6.3.2	Vytváranie profilov otázok a používateľov	34
6.3.3	Zostavenie zoznamu odporúčaní	36
6.4	Charakteristika dát	37
7	Experimentálne overenie	39
7.1	Metodológia experimentu	40
7.2	Metriky hodnotenia výsledkov	41
7.3	Vyhodnotenie výsledkov experimentu	42
7.3.1	Aktivita používateľov	42
7.3.2	Vyhodnotenie sledovaných metrík	43
7.3.3	Dotazník odoberateľov	44
7.3.4	Diskusia	45
8	Zhodnotenie	46
	Literatúra	47
A	Plán práce na diplomovom projekte	A-1
A.1	Plán práce na diplomovom projekte I	A-1
A.2	Plán práce na diplomovom projekte II	A-2
A.3	Plán práce na diplomovom projekte III	A-3
B	Technická dokumentácia systému	B-1
B.1	Modul pre registráciu nových používateľov	B-1
B.2	Modul pre generovanie a odosielanie informačných bulletinov	B-1
B.3	Modul pre spracovanie statickej zálohy databázy	B-3
B.4	Databázový model	B-4
C	Výsledky dotazníku odoberateľov	C-1

D Príspevok na konferencii IIT.SRC 2018

D-1

E Elektronické médium

E-1

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 komúnít, ktoré produkujú veľké množstvo používateľmi vytváraného obsahu. Medzi populárne druhy takýchto online komúnít 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 a to napriek tomu, že existujúce informačné bulletiny nespĺňajú očakávania komunity a sú pre ne charakteristické viaceré problémy, napr. akým spôsobom ich vhodne personalizovať a ako zároveň zabezpečiť diverzitu ich obsahu.

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 odpovedanie na otázky. V našej práci sa zameriavame predovšetkým na skúmanie vplyvu zavedenia diverzity a podpory aktuálnosti na úspešnosť personalizovaného odporúčania otázok vo forme informačných bulletinov.

TODO – aktualizovať a doplniť štruktúru práce.

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 riešiť informačné bulletiny.

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

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 otáznym vzhľadom na veľkú diverzitu obsahu aj používateľov.

Automaticky generované personalizované 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 otvorili 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 klikov (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 Diskusia

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.

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 a odporúčanie

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 komunit, od všeobecnejších, ako je napr. komunita venujúca sa matematike¹, 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.

Tematicky zamerané CQA systémy majú väčší potenciál pre vznik dlhodobu 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 vysporiadať 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 zvyrazňuje 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 2017⁴ 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>

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

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

3.3 Odporúčacie systémy

Odporúčacie systémy sú softvérové nástroje a techniky ktoré používateľom ponúkajú položky, ktoré by pre nich mohli byť nejakým spôsobom zaujímavé alebo užitočné [9]. 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 *filtrovaní založeného na obsahu* (angl. *Content-based filtering*) [10]. Tieto dva prístupy môžu byť tiež kombinované v hybridných odporúčacích systémoch.

3.3.1 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 [10]. 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 prostredníctvom oceňovania parametrov štatistických modelov pre používateľské hodnotenia.

3.3.2 Filtrovanie na základe obsahu

Odporúčanie čisto prostredníctvom 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 [9]. Odporúčania v takýchto prístupoch vznikajú na základe prekryvu týchto dvoch modelov.

3.4 Odporúčanie v CQA systémoch

V kontexte CQA systémov je problematika odporúčania a odporúčacích systémov častým objektom výskumu [11].

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 [11].

3.4.1 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. 3.3.1) alebo filtrovania na základe obsahu (kap. 3.3.2).

Forma prezentovania odporúčaných otázok sa tiež môže líšiť. Časté je napríklad zobrazenie otázok, ktoré by používateľa mohli 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.

3.4.2 Smerovanie otázok

Na rozdiel od pomerne štandardného odporúčania otázok, v prípade smerovania otázok (angl. *Question routing*) je prístup k odporúčaniam 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].

3.4.3 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 systémov 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 spoločné 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.

3.4.4 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 používa-

fskej základne tvoria používatelia, ktorí sú noví, alebo nemajú žiadnu aktivitu. Pre odporúčanie otázok (kapitola 3.4.1) je tak problém zostaviť profil používateľa, na základe ktorého sa vykonáva mapovanie na model relevantného obsahu. V prípade smerovania otázok (kapitola 3.4.2) je zase problém zostaviť model reprezentujúci expertízu daného používateľa.

3.4.5 Problém filtračnej bubliny

Ďalším problémom, ktorému sa však v oblasti odporúčania obsahu CQA systémov venuje menej pozornosti [11], 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.

3.5 Informačné bulletiny v CQA systémoch

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

Súčasný výskum v oblasti CQA systémov [11] 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.

3.5.1 Informačné bulletin v sieti Stack Exchange

Sieť Stack Exchange⁵, 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 bulletin komunit 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 spokojní⁶. Medzi problémy, ktoré najčastejšie používatelia vytýkajú súčasnému informačnému bulletinu patria:

- **Náhodne generovaný obsah** – Sekcia nezodpovedaných otázok obsahuje náhodný výber otázok bez odpovedí. Pravdepodobnosť, že používateľ vie na niektorú z nich odpovedať, je tak veľmi malá⁷.
- **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 obsahu⁸.
- **Malá rôznorodosť obsahu** – Hlavne pokročilejší a aktívnejší používatelia by chceli v informačnom bulletin vidieť okrem rôznych otázok aj iný obsah, okrem iného napr. rôzne štatistiky aktivity komunity, zoznam ocenených používateľov alebo príbuzný obsah z komunitných blogov⁹.
- **Aktuálnosť obsahu** – Informačný bulletin niekedy obsahuje veľmi staré otázky, ktoré už nie sú relevantné¹⁰.

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

⁵ <https://stackexchange.com>

⁶ <https://meta.stackexchange.com/q/247298>. Prevzaté 31.4.2017.


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>

v prípade väčších komunit, akou je napríklad Stack Overflow¹¹, často obsahuje otázky, ktoré nie sú z oblastí záujmu používateľa.



 **stackoverflow** newsletter

Top new questions this week:

[What is python .. \("dot dot"\) notation syntax?](#)

I recently came across a syntax I never seen before when I learned python nor in most tutorials, the .. notation, it looks something like this: `f = 1..__truediv__ #` or `1..__div__` for python 2 ...



pythonpython-3.xsyntaxoperatorspython-2.x

 asked by [abccd](#) 120 votes
 answered by [Paul Rooney](#) 136 votes

[Why is -1**2 a syntax error in JavaScript?](#)

Executing it in the browser console it says `SyntaxError: Unexpected token **`. Trying it in node: `> -1**2` ...
...^C I thought this is an arithmetic expression where `**` is the power ...

javascriptexponentiationecmascript-2016



 asked by [psmith](#) 33 votes
 answered by [torazaburo](#) 49 votes

Greatest hits from previous weeks:

[How to undo last commit\(s\) in Git?](#)

I committed the wrong files to Git. How can I undo that commit?



gitgit-rebasegit-commitgit-resetgit-revert

 asked by [Hamza Yerlikaya](#) 13798 votes
 answered by [Esko Luontola](#) 14309 votes

[Setting "checked" for a checkbox with jQuery?](#)

I'd like to do something like this to tick a checkbox using jQuery: `$(".myCheckBox").checked(true)`; or `$(".myCheckBox").selected(true)`; Does such a thing exist?

javascriptjquerycheckboxselectedchecked


 asked by [tpower](#) 3017 votes
 answered by [Xian](#) 4667 votes

Can you answer these?

[C++ using global variable shows 100% slower than a pointer, when using pthread?](#)

I've got a quite program to show the performance of 2 similar programs, both uses 2 threads to do calculation. The core difference is that one uses a global variable, another uses a "new" object, as ...

c++linuxperformancevariables pthreads

 asked by [Troskys](#) 5 votes

[Why does Visual Studio compiler allow violation of private inheritance in this example?](#)

I found very strange behavior of `std::unique_ptr` in visual studio both 2013 and 2017. Let's consider an example: `class Base { public: virtual ~Base() = default; virtual void Foo() = 0; }; ...`

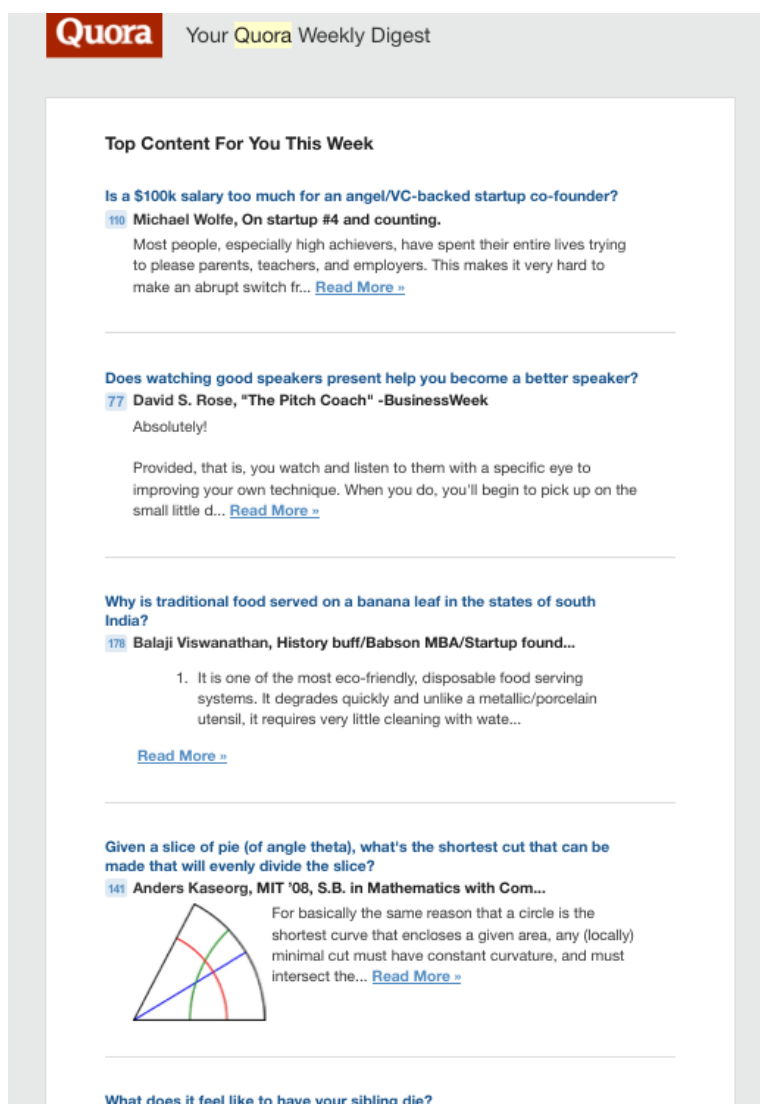
Obr. 3.1: Informačný bulletin komunity Stack Overflow, 25. apríl 2017. Adresát tohto bulletinu má aktivitu prevažne v PHP a SQL, čo vôbec nezodpovedá obsahu vygenerovaného bulletinu.

¹¹<https://stackoverflow.com>

3.5.2 Informačný bulletin portálu Quora

Quora¹² 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 zvlášť¹³ (Obr. 3.2). Nie je však známe, akým spôsobom je vyberaná personalizovaná časť informačného bulletinu.



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

¹²<https://quora.com>

¹³<http://businessinsider.com/quora-emails-2012-8>

3.5.3 CQA systémy bez informačných bulletinov

Viacere 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>

4 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 [9]. Dôvodom je práve náchylnosť takéhoto odporúčania na vyvolanie problému filtračnej bubliny (viď. kapitola 3.4.5).

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

4.1 Diverzita v odporúčacích systémoch

Tématická diverzifikácia je metóda napomáhajúca vyváženosti a diverzite personalizovaného odporúčania s cieľom lepšie reflektovať kompletne 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 3.3.1). Vnútorná miera podobnosti položiek zoznamov bola určovaná prostredníctvom metriky založenej na taxonomickej klasifikácii 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 diverzity 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 $|S| = k$, $score(S) \geq \theta$ 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] nezaobrá priamo diverzitou v odporúčacích systémoch, sú paralely s touto oblasťou výrazné. Motiváciou pre takýto 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.

4.2 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 [9].

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 vyvíjajúcej 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 problémovjšie 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 črt, ktoré nie sú ovplyvnené časom (napr. nadpis, text alebo autor obsahu, na rozdiel od počtu hlasov alebo dátumu uverejnenia), alebo tiež upravenie hodnôt týchto črt vzhľadom na relatívnu aktuálnosť obsahu.

4.3 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 [11].

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 diverzifikácie 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 úspešné 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%.

4.4 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.

5 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 v rámci platformy Stack Exchange, ktorá patrí medzi najpopulárnejšie CQA systémy v súčasnosti a tvorí ju viac ako 160 komunít zameraných na rôzne oblasti.

Informačný bulletin tvorí viacero sekcií, ktoré vyžadujú rôzne spôsoby odporúčania. Konkrétne ide o odporúčanie nových otázok a odporúčanie vyriešených otázok, pričom v prvom prípade je dôležité pozerať sa na expertízu používateľa a v druhom prípade na oblasti jeho záujmu.

Hypotéza 1

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.

Hypotéza 2

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.

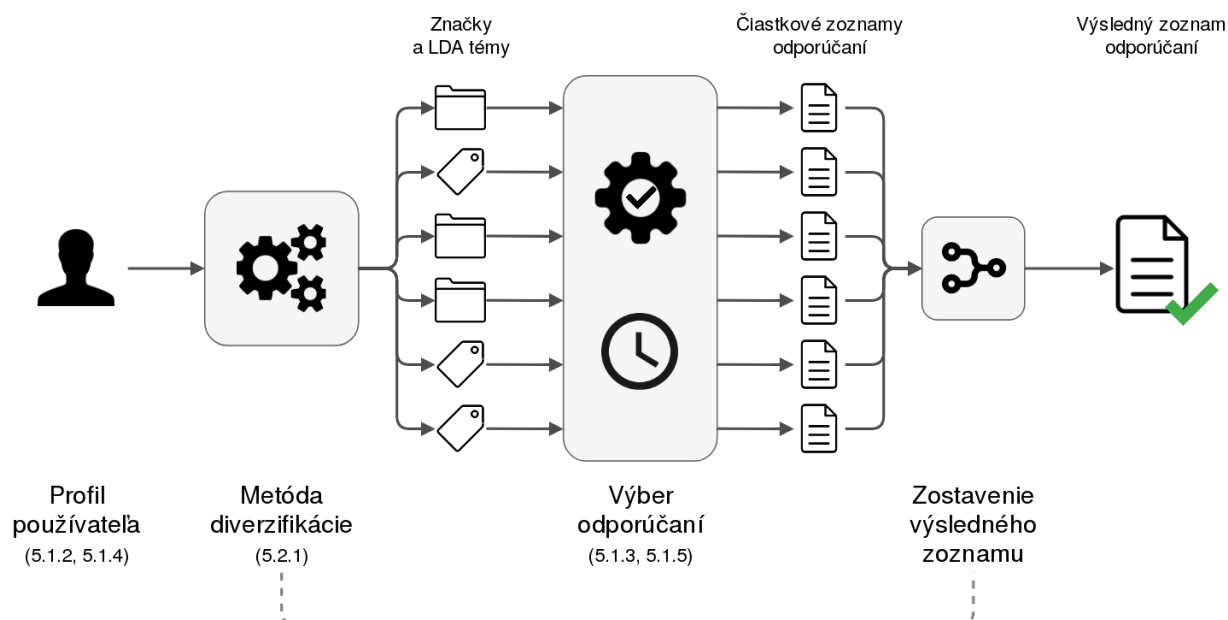
Prehľad navrhovanej metódy

Na začiatok uvádzame stručný súhrn navrhutej metódy zostavovania personalizovaných informačných bulletinov. Celkový pohľad na metódu ilustruje obrázok 5.1, Jednotlivé časti metódy sú podrobne opísané v nasledujúcich kapitolách.

- Pre zostavenie personalizovaných informačných bulletinov využijeme odporúčaciu metódu filtrovania na základe obsahu (kap. 3.3.2), keďže je v prípade malej používateľskej aktivity efektívnejšia ako kolaboratívne filtrovanie (viď kap. 3.4.4).
- Diverzifikáciu odporúčaného obsahu zabezpečujeme na úrovni tématického pre-filteru pred samotným procesom odporúčania – pre každú nezávislú tému sa vytvára samostatný zoznam odporúčaní. Pri návrhu tejto metódy sme vychádzali z [25].
- Výsledný zoznam odporúčaného obsahu vzniká spojením jednotlivých zoznamov z kroku

diverzifikácie.

- Pri odporúčaní sa okrem diverzity a relevancie obsahu uvažuje aj jeho aktuálnosť.
- Navrhnutá metóda reflektuje implicitnú aj explicitnú spätnú väzbu používateľa vo forme interakcie s informačným bulletinom a profil používateľa sa na jej základe aktualizuje.



Obr. 5.1: Schéma metódy zostavovania personalizovaných informačných bulletinov.

5.1 Návrh metódy personalizovaného odporúčania

Pri vytváraní personalizovaného informačného bulletinu sa zameriavame na odporúčanie relevantných otázok jednotlivým používateľom CQA systému prostredníctvom aplikovania metódy filtrovania na základe obsahu (kapitola 3.3.2). Pre účely filtrovania na základe obsahu je potrebné definovať a zostaviť profily reprezentujúce jednak otázky, a tiež používateľov CQA systému.

5.1.1 Profil otázok

Profil otázky sa skladá z troch nezávislých modelov, ktoré sa na samotnú otázku pozerajú z rôznych perspektív.

1. Kategrický model otázok

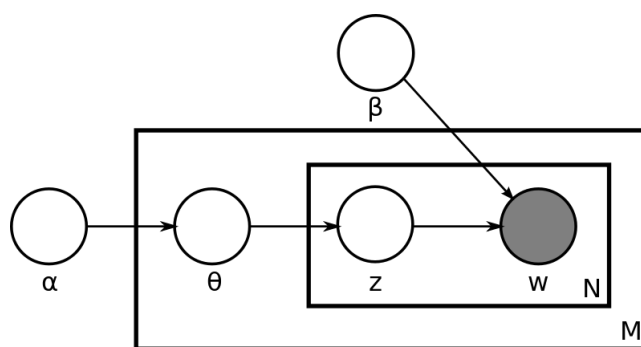
Tento model reprezentuje otázku na najvyššej úrovni ako prislúchajúcu do určitých kategórií. Kategórie otázok sú v rámci platformy Stack Exchange reprezentované ako značky (angl. *tags*). Každá otázka môže obsahovať viacero značiek.

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

2. Tématický model otázok

Tématický model využíva metódu latentnej Dirichletovej alokácie (angl. *Latent Dirichlet Allocation - LDA*) [28] na určenie latentných tém, ktorým sa daná otázka venuje.

LDA vektor tohto modelu reprezentuje distribúciu otázky v rámci jednotlivých latentných tém. Samotné LDA témy sa odvodzujú z nadpisu a textu otázky. Pre optimalizáciu modelu a zanedbanie tém s veľmi nízkou distribúciou sa do úvahy berú len latentné témy tvoriace 75% z celkovej distribúcie, teda tretí kvartil. Jednotlivé hodnoty tém sú následne normalizované, aby tvorili 100%.



Obr. 5.2: Platňová notácia LDA modelu.¹

Nastavenie LDA

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

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

M – Celkový počet dokumentov.

N – Celkový počet slov vo všetkých dokumentoch.

Z – N -rozmerný vektor identít tém všetkých slov vo všetkých dokumentoch.

W – N -rozmerný vektor identít všetkých slov vo všetkých dokumentoch.

n – Celkový počet tém.

α – Úvodná váha tém v dokumente. $1/n$ zabezpečí v úvode rovnomernú distribúciu.

β – Úvodná váha slov v téme. $1/n$ zabezpečí v úvode rovnomernú distribúciu.

¹Prevzaté 10.12.2017 z https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

θ_d – Distribúcia tém v dokumente d .

κ – Úpadok učenia; parameter kontrolujúci mieru učenia v online metóde učenia.

τ_0 – Posun učenia; znižuje váhu počiatočným iteráciám v online metóde učenia.

Nakoľko jednotlivé komunity v rámci platformy Stack Exchange sú pomerne úzko zamerané, predpokladáme, že je dostačujúce natrénovať LDA model na vzorke archívnych dát a nie je 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 je zostavený z nadpisu a textu samotnej otázky. Pred výpočtom je text lematizovaný a sú z neho odstránené stop slová.

$$\text{tf-idf}(t, d) = \text{tf}(t, d) \times \log \frac{n_d}{1 + \text{df}(d, t)}$$

5.1.2 Profil používateľov

Pre každého používateľa uvažujeme dva nezávislé podprofily – jeden profiluje záujem používateľa o určité témy a otázky, druhý profiluje jeho expertízu v určitej oblasti. podprofil reflektujúci záujem používateľa je použitý pri odporúčaní otázok, ktoré by používateľa mohli zaujímať, druhý pri odporúčaní otázok, ktoré by mohol vedieť zodpovedať. Oba podprofily sú z pohľadu svojej štruktúry presne rovnaké.

Takýto prístup k rozdeleniu profilu používateľa podľa záujmu a expertízy je vo svojej podstate unikátny, nakoľko ho využíva iba práca [30]. V ostatných prácach autori buď modelujú len jeden z týchto aspektov, alebo ich vôbec nerozlišujú [11].

Profil používateľa je zostavený na základe jeho aktivity a analogicky k profilu otázok sa skladá z troch vektorov:

1. Prvý vektor reprezentuje aktivitu používateľa naprieč značkami – každý rozmer predstavuje jednu značku, v ktorej má používateľ aktivitu. Hodnoty v jednotlivých rozmeroch predstavujú podiel aktivity v danej značke voči celkovému množstvu aktivity používateľa.
2. Druhý vektor analogicky k prvému reprezentuje aktivitu používateľa naprieč LDA témami, v ktorých má používateľ aktivitu.

3. Tretí vektor reprezentuje zastúpenie jednotlivých výrazov v textoch otázok, v ktorých má používateľ aktivitu.

Záujmový podprofil

Podprofil predstavujúci záujem používateľa je zostavený zo všetkých otázok, ktoré používateľ položil, alebo ktoré označil za obľúbené. Každá takáto otázka do podprofilu prispieva rovnakou váhou.

Expertízny podprofil

Podprofil modelujúci expertízu používateľa sa skladá z otázok, ktoré používateľ označil za obľúbené alebo na ktoré používateľ odpovedal, pričom ich dopad na podprofil expertízy je závislý od skóre jeho odpovede. Nezáporné skóre prispieva do podprofilu pozitívne – signalizuje to teda fakt, že používateľ danej téme rozumie. Odpoveď so záporným skóre naopak signalizuje, že používateľ danej téme nerozumie, čo je reflektované aj v jeho expertíznom podprofile. Odpovede označené za akceptované do podprofilu prispievajú s koeficientom 1.75, čo je priemerný počet odpovedí na otázky s akceptovanou odpoveďou.

Komentáre

Okrem pokladania otázok, odpovedania alebo označenia za obľúbené uvažujeme aj používateľove komentáre. Keďže však zo samotného faktu, že používateľ niečo okomentoval nemožno jednoznačne určiť, či táto aktivita predstavuje jeho záujem alebo expertízu, prispievajú otázky, ktoré používateľ okomentoval do celého profilu – záujmového aj expertízneho, avšak s koeficientom 0.3. Tento koeficient vychádza z faktu, že v rámci platformy Stack Exchange pripadajú na jednu komentovanú otázku alebo odpoveď priemerne tri komentáre.

Spätná väzba

Do oboch podprofilov prispievajú využitím rovnakého princípu aj implicitná a explicitná spätná väzba používateľa – teda jeho kliknutie na konkrétnu otázku v informačnom bulletine, alebo jej explicitné kladné alebo záporné ohodnotenie. Implicitná spätná väzba prispieva do podprofilov s koeficientom 0.3, explicitná s koeficientom ± 1 .

Vzorce pre zostavenie záujmového a expertízneho podprofilu používateľa

$$\mathbf{UP}_{\text{interest}} = \sum_{q \in Qa, Qf} T_q + \sum_{q \in Qc, Qif} 0.3 \times T_q + \sum_{q \in Qef} \pm 1 \times T_q$$

$$\mathbf{UP}_{\text{expertise}} = \sum_{q \in Qw, Qf} T_q + \sum_{q \in Qac} 1.75 \times T_q + \sum_{q \in Qc, Qif} 0.3 \times T_q + \sum_{q \in Qef} \pm 1 \times T_q$$

T_q – profil otázky q .

Qa – množina položených otázok

Qf – množina otázok označených za obľúbené

Q_c – množina komentovaných otázok

Q_w – množina zodpovedaných otázok

Q_{ac} – množina zodpovedaných otázok s označením akceptovanej odpovede

Q_{if} – množina otázok s implicitnou spätnou väzbou

Q_{ef} – množina otázok s explicitnou spätnou väzbou

5.1.3 Výber odporúčaného obsahu

Pre účely zostavovania zoznamu odporúčaných otázok používame prístup analogický štandardným nástrojom pre vyhľadávanie informácií (angl. *Information Retrieval Engines*). V našom prípade sú dokumentmi samotné otázky a dopytom je príslušný profil používateľa. Pre ohodnocovanie podobnosti profilov je použitý skalárny súčin vektorov reprezentujúcich profily otázok a používateľa.

Skalárny súčin (angl. *dot product*) vektorov \mathbf{a} a \mathbf{b} .

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^n a_i b_i = a_1 b_1 + a_2 b_2 + \dots + a_n b_n$$

5.1.4 Riešenie problému studeného štartu

Pre eliminovanie problému studeného štartu (kapitola 3.4.4) z pohľadu prvotného odporúčania otázok využívame offline natrénovanie našich metód na archívnych dátach platformy Stack Exchange.

V prípade nových používateľov, ktorí v systéme nemajú žiadnu, alebo iba nedostatočnú aktivitu, sa zo začiatku namiesto profilu konkrétneho používateľa používa komunitný profil, ktorý je zostavený analogicky k profilu používateľa, no pozostáva z celkovej aktivity v systéme.

5.1.5 Aktuálnosť odporúčaní

Pri zostavovaní odporúčaní do informačného bulletinu sa uvažujú len otázky, ktoré pochádzajú z obdobia od posledného zostavenia informačného bulletinu.

Aby profil používateľa reflektoval meniace sa záujmy používateľa v čase, zaviedli sme pri

zostavovaní profilu z používateľovej aktivity exponenciálny *faktor úpadku* (angl. *decay factor*):

$$y_{n+t} = y_n(1 - d)^t$$

y_n – aktuálna váha danej hodnoty vo vektore profilu používateľa

y_{n+t} – váha danej hodnoty po aplikovaní faktoru úpadku

t – čas od aktualizácie profilu používateľa – počet aktualizácií profilu.

d – percentuálny pokles danej hodnoty; vypočíta sa ako podiel množstva používateľovej aktivity za čas t a jeho celkového množstva aktivity.

5.2 Návrh metódy diverzifikácie odporúčaní

Diverzifikácia obsahu personalizovaného informačného bulletinu sa vykonáva ešte pred samotným zostavovaním zoznamu odporúčaní formou pre-filteru. Diverzifikácia spočíva vo výbere *značiek* a *tém*, pričom následne v kroku zostavovania zoznamov odporúčaní sa pre každý zoznam uvažuje len obsah prislúchajúci do danej značky alebo témy. Okrem výberu značiek a tém je 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 diverzifikácii by v praxi nebol realizovateľný. Ako príklad môžeme uviesť systém Stack Overflow, kde denne pribudne cca 7000 nových otázok, čo by v prípade generovania týždenného informačného bulletinu znamenalo vypočítať podobnosť až cca 50000 profilov otázok s profilom každého používateľa.

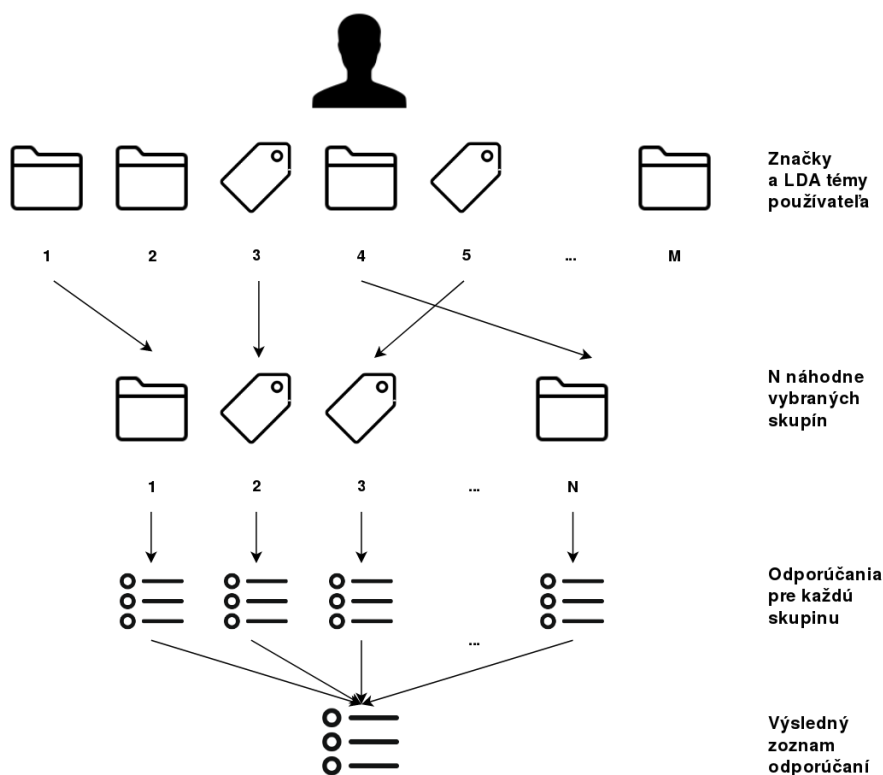
Cieľom našej práce je vyhodnotiť dopad diverzifikácie odporúčaní na personalizované informačné bulletiny CQA systémov. Za týmto účelom sme navrhli metódu tématického vzorkovania (angl. *Thematic sampling*).

Metóda tématického vzorkovania

Pri návrhu tejto metódy sme sa inšpirovali rovnomennou metódou z [25], ktorú sme prispôbili podmienkam špecifickým pre našu prácu. Proces tématického vzorkovania ilustruje schéma na obrázku 5.3.

1. Pre každého používateľa vyberáme náhodne z jeho profilu $\left\lceil \frac{n}{2} \right\rceil$ značiek a $\left\lfloor \frac{n}{2} \right\rfloor$ tém, kde n je počet položiek vo výslednom zozname. Pri výbere sa obmedzujeme na značky a témy, ktoré spadajú do štvrtého kvartilu, najmenej však vyberáme z piatich. Potláčame tak značky a témy, ktoré majú pre používateľa nízku relevanciu.

2. Prostredníctvom vyššie opísanej metódy (kapitola 5.1.3) sa pre každú takúto značku a tému následne zostavuje zoznam n odporúčaní, pričom sa uvažujú len otázky prislúchajúce tejto téme alebo značke.
3. Následne náhodne vyberáme vzorky z jednotlivých zoznamov do výsledného zoznamu odporúčaní, pričom pravdepodobnosť výberu otázky z určitého zoznamu je proporčná k relevancii tohto zoznamu pre daného používateľa.
4. Konkrétne otázky do výsledného zoznamu odporúčaní z jednotlivých zoznamov sa vyberajú z $m \times n$ najrelevantnejších otázok daného zoznamu, pričom m je relatívna relevancia tohto zoznamu.



Obr. 5.3: Schéma metódy diverzifikácie odporúčaní tématickým vzorkovaním.

Pre použitie práve takto navrhnutej metódy diverzifikácie sme sa rozhodli z viacerých dôvodov. Dôležitým faktorom bola jej veľmi dobrá škálovateľnosť, nakoľko umožňuje už v prvom kroku obmedziť množinu otázok, z ktorých sa vyberajú odporúčania. To je veľmi dôležité najmä v prípade systémov s veľkým množstvom obsahu, akými je aj platforma Stack Exchange. V súvislosti s možnosťou škálovania nám tiež z preštudovaných metód diverzifikácie príde metóda tématického vzorkovania najprirodzenejším prístupom k diverzifikácii odporúčaní. V neposlednom rade zavážil aj fakt, že pôvodná metóda, ktorú je naša metóda inšpirovaná, bola v minulosti úspešne otestovaná v nekontrolovanom online experimente v CQA systéme Yahoo! Answers [25].

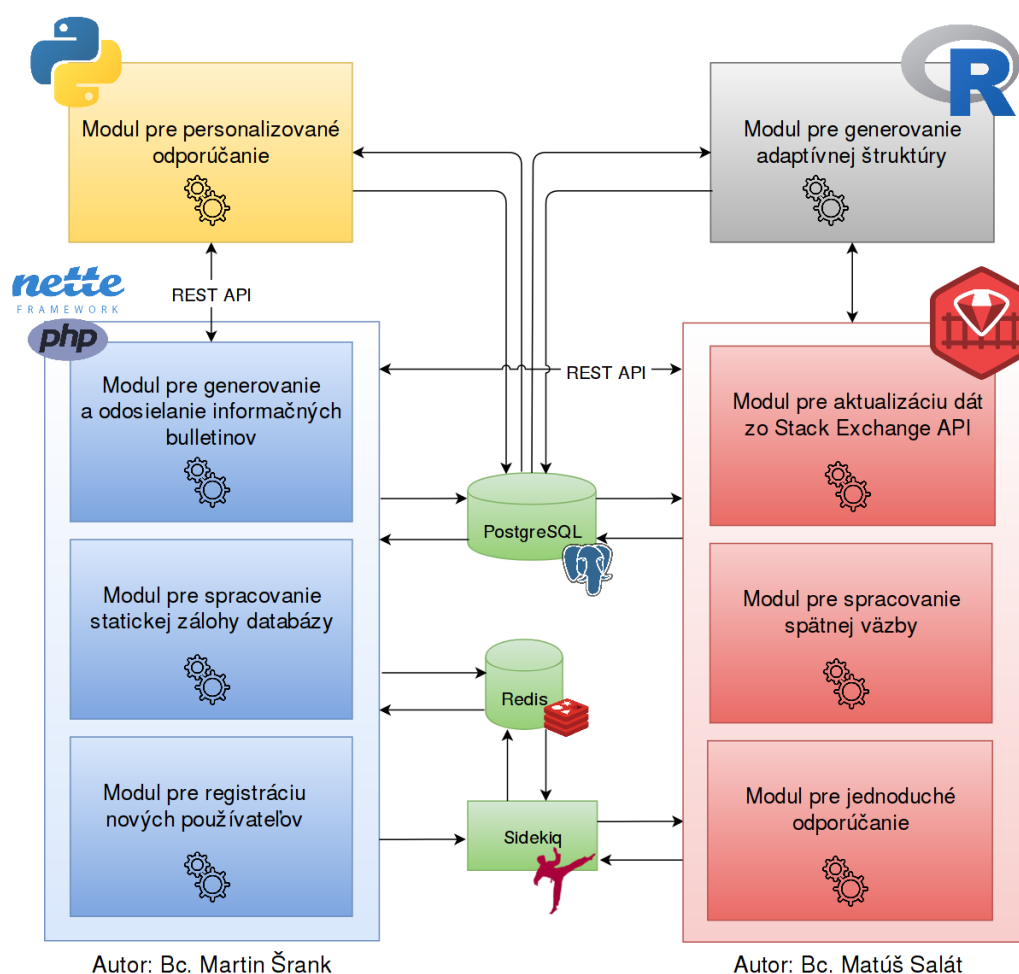
5.3 Zhrnutie

Navrhli sme metódu personalizovaného odporúčania v informačných bulletinoch, ktorá kladie dôraz na diverzifikáciu a aktuálnosť odporúčaného obsahu. Už samotné zameranie sa práve na informačné bulletiny je v oblasti CQA systémov v podstate unikátne. Na rozdiel od väčšiny výskumu v oblasti odporúčania v CQA systémoch naša metóda využíva princíp odporúčania otázok, a nie ich smerovania.

V profile používateľa rozlišujeme jeho záujmy od jeho expertízy, a odporúčame mu relevantný obsah podľa toho, o akú sekciu informačného bulletinu sa jedná. Používateľský profil sa časom vyvíja tak, aby vždy reflektoval aktuálne zameranie používateľa, prostredníctvom exponenciálneho úpadkového faktoru. Ako riešenie problému studeného štartu sme navrhli tzv. komunitný profil, ktorý reprezentuje agregované záujmy a expertízu všetkých používateľov CQA systému.

6 StackLetter – personalizovaný informačný bulletin pre Stack Exchange

StackLetter je systém pre vytváranie a rozosielanie personalizovaných informačných bulletinov v rámci platformy Stack Exchange. Tento systém vznikol ako súčasť spolupráce autora tejto práce a Bc. Matúša Saláta [31] v rámci dvoch diplomových prác riešených v akademickom roku 2016/17 a 2017/18. Celý systém sa skladá z viacerých spolupracujúcich modulov, ktoré však boli vyvíjané samostatne a sú od seba navzájom nezávislé. Rozdelenie systému na nezávislé moduly umožňuje rýchlejší vývoj, ako aj možnosť jednoduchého rozširovania systému v budúcnosti. Architektonický prehľad celého systému znázorňuje obrázok 6.1. Ďalej v tejto práci opisujem len mnou navrhnuté moduly.



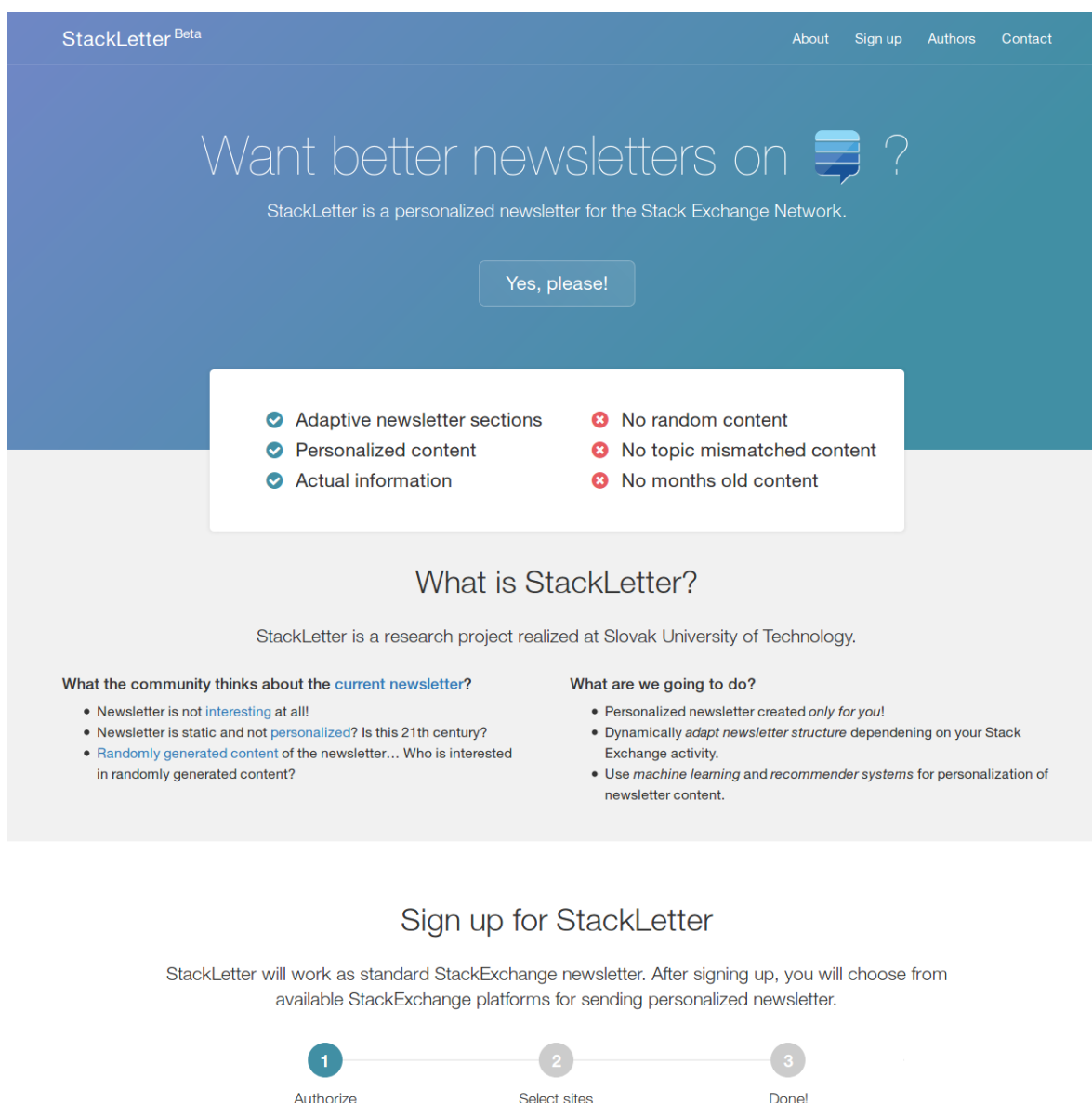
Obr. 6.1: Architektonický prehľad systému StackLetter.

6.1 Prehľad modulov

Modul pre registráciu nových používateľov

Tento modul predstavuje používateľmi viditeľnú časť systému. Prostredníctvom webovej stránky systému – `www.stackletter.com` – sa používatelia môžu prihlásiť k odoberaniu informačného bulletinu pre niektoré z ponúkaných komunít platformy Stack Exchange, ako aj spravovať svoje nastavenia týkajúce sa odosielania informačných bulletinov.

Používatelia sa prihlasujú prostredníctvom svojho Stack Exchange konta využitím protokolu OAuth.



Obr. 6.2: Snímka registračnej stránky StackLetter.com.

Modul pre generovanie a odosielanie informačných bulletinov

Tento modul je zodpovedný za samotné zostavovanie a odosielanie informačných bulletinov jednotlivým zaregistrovaným používateľom. Prostredníctvom REST API komunikuje s modulmi pre zostavovanie odporúčaní a štruktúry informačných bulletinov a na základe ich odpovedí vygeneruje naformátované informačné bulletiny, ktoré sú následne rozosielené prostredníctvom služby SendGrid¹.

Modul pre spracovanie statickej zálohy databázy

Modul bol navrhnutý a implementovaný za účelom rýchleho a efektívneho importovania počiatočných dát platformy Stack Exchange dostupných vo forme XML exportu do našej internej reprezentácie v databáze PostgreSQL.

Modul pre personalizované odporúčanie

Modul zabezpečuje samotné vytváranie personalizovaných zoznamov odporúčaní prostredníctvom metódy navrhutej v tejto práci. Podrobne sa realizácii tohto modulu venujeme v kapitole 6.3.

Podrobný technický a architektonický popis jednotlivých modulov sa nachádza v prílohe B – *Technická dokumentácia systému StackLetter*.

6.2 Použité technológie a služby

- Jednotlivé moduly zabezpečujúce infraštruktúru systému StackLetter sú implementované v jazyku PHP² verzie 7.1 s použitím MVC frameworku Nette 2.4³.
- Systém na ukladanie dát využíva relačný databázový systém PostgreSQL vo verzii 9.6 a vnútropamäťové dátové úložisko Redis vo verzii 4.0.
- Systém komunikuje s platformou Stack Exchange prostredníctvom Stack Exchange API v2.2 a vykonáva autentifikáciu používateľov prostredníctvom protokolu OAuth 2.0.
- Na hromadné odosielanie informačných bulletinov používateľom prostredníctvom e-mailu je využitá služba SendGrid.
- Modul pre zostavovanie personalizovaných informačných bulletinov so zameraním na diverzitu je implementovaný v jazyku Python 3.6 s použitím knižníc scikit-learn, numpy (pre prácu s modelmi) a Flask (pre implementáciu REST API).

¹<https://sendgrid.com>

²<https://php.net>

³<https://nette.org>

6.3 Realizácia personalizovaného odporúčania

Modul pre personalizované odporúčanie je navrhnutý ako samostatný modul, ktorý poskytuje REST API pre komunikáciu s modulom pre generovanie a odosielanie informačných bulletinov. Vďaka tomu je možné do budúcnosti jednoducho nahradiť tento modul iným modulom bez potreby výrazných zásahov do existujúcich častí systému.

Každý informačný bulletin sa skladá z troch sekcií, ktoré sú vyberané personalizovane pre jednotlivých používateľov. Zoznam všetkých dostupných sekcií a aj výber sekcií pre konkrétneho používateľa je témou práce spolužiaka Matúša Saláta[31], preto sa mu v tejto práci nevenujeme.

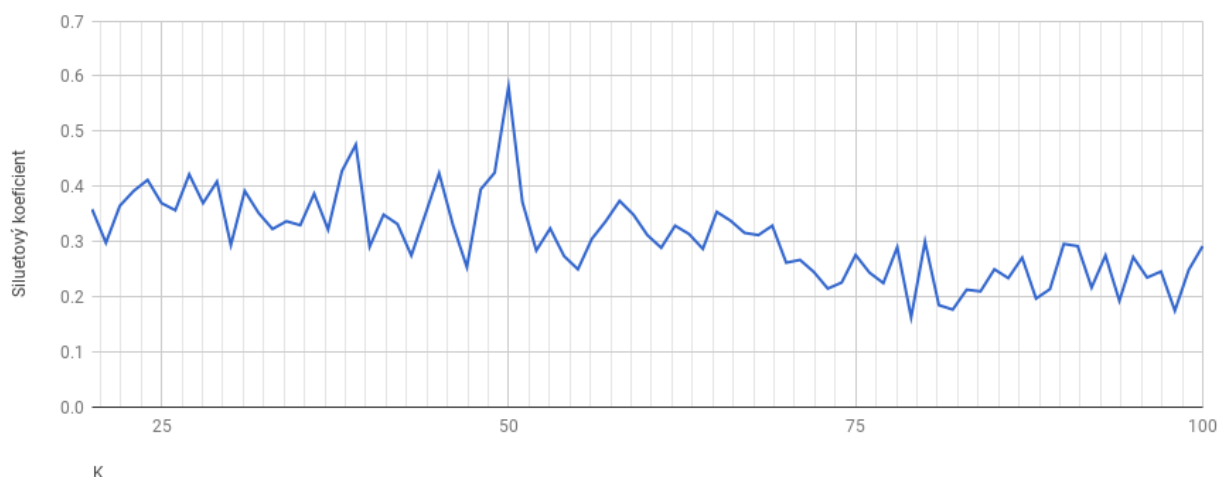
6.3.1 Zostavenie LDA a TF-IDF modelov

Slovník

Pre LDA a TF-IDF modely využívané v navrhutej metóde sme vytvorili slovnú zásobu pozostávajúcu z 500 tisíc náhodne vybraných otázok z komunity Stack Overflow. Texty prešli predspracovaním v podobe odstránenia formátovania (HTML tagov), diakritiky, anglických stop slov a lematizáciou. Minimálna dokumentová frekvencia bola stanovená na 5 a maximálna dokumentová frekvencia na 99%. Okrem toho bola maximálna veľkosť slovníka stanovená na 150 tisíc výrazov, pričom skutočný počet výrazov bol cca 152 tisíc. Na vytvorenie slovníku bol použitý *CountVectorizer* z knižnice *scikit-learn*.

Nastavenie a trénovanie LDA modelu

Pravdepodobne najdôležitejším parametrom pri latentnej Dirichletovej alokácii je určenie počtu LDA tém. Pre určenie počtu sme na testovacie dáta (náhodne vybraných stotisíc otázok) opakovane aplikovali zhľukovanie prostredníctvom metódy K-Means, pričom hodnotu k (počet zhľukov) sme vyberali z intervalu $< 20, 100 >$. Pre každú hodnotu k sme následne vypočítali priemerný siluetový koeficient (viď. Obr. 6.3). Na základe tejto analýzy sme stanovili počet LDA tém na 50.



Obr. 6.3: Siluetový koeficient vzhľadom na K – počet zhlukov.

LDA model sme natrénovali na vzorke stotisíc náhodne zvolených otázok zo Stack Overflow, pričom sme využili slovník výrazov vytvorený v predošlom kroku. Ostatné parametre modelu boli nastavené tak, ako boli definované v kapitole 5.1.1. Bola použitá implementácia LDA z knižnice *scikit-learn*.

TF-IDF model

TF-IDF model zostavujeme zvlášť pre každého používateľa z otázok, s ktorými interagoval, a to pri každom pretrénovaní používateľských profilov, ktoré sa vykonáva raz denne pre používateľov s denným informačným bulletinom a raz týždenne pre používateľov s týždenným bulletinom. Využíva sa implementácia *TfidfTransformer* z knižnice *scikit-learn*.

6.3.2 Vytváranie profilov otázok a používateľov

Profil otázky

Profil otázky sa skladá z troch vektorov: 1) vektor značiek (tagov), ktorými je otázka označená, 2) vektor reprezentujúci distribúciu otázky v rámci LDA tém a 3) TF vektor. Vektor značiek sa vytvára za behu priamo požiadavkou do databázy. TF vektor sa tiež vytvára až za behu z textu otázky, nakoľko sa jedná o veľmi rýchlu operáciu. Vektor distribúcie LDA tém sa vytvára vždy pri vzniku novej otázky, alebo po interakcii používateľa s otázkou, pre ktorú ešte nebol vytvorený. Tento vektor sa po vytvorení ukladá do databázy. Vytváranie profilov pre novo vzniknuté otázky sa vykonáva raz denne.

Profil používateľa

Profil používateľa sa nevytvára za behu, ale sa perzistuje v binarizovanej podobe do súboru,

odkiaľ sa v prípade použitia načítava. Profil používateľa sa delí na podprofily záujmu a expertízy, pričom tie sa vytvárajú z profilov otázok, s ktorými používateľ interagoval.

Každý podprofil sa skladá z vektoru značiek, ktorý predstavuje relatívne zastúpenie značiek v otázkach, s ktorými používateľ interagoval, normalizovaných do intervalu $< 0, 1 >$; vektoru LDA tém, ktorý je vytvorený analogicky a tiež normalizovaný do intervalu $< 0, 1 >$, a matice TF vektorov z profilov príslušných otázok. Z tejto matice sa následne predpočíta aj matica TF-IDF.

Aktualizácia profilov používateľov

Profil používateľa sa aktualizuje v pravidelných intervaloch na základe jeho preferencie prijímania informačných bulletinov, teda denne alebo týždenne. Pre zachytenie meniacich sa záujmov používateľa v čase (viď kapitola 5.1.5) sa využíva exponenciálny faktor úpadku: $df = (1 - d)^t$.

Čas (t) sa udáva ako počet aktualizácií profilu, a percentuálny pokles (d) sa vypočíta ako podiel množstva používateľovej aktivity od poslednej aktualizácie profilu a celkového množstva používateľovej aktivity. Množstvo používateľovej aktivity sa počíta s príslušnými váhami za jednotlivé aktivity (viď kapitola 5.1.2). Faktor úpadku sa počíta zvlášť pre oba používateľské podprofily.

Pri aktualizácii profilu používateľa sa najprv týmto faktorom úpadku prenasobia vektory reprezentujúce doterajšiu aktivitu a následne sa k nim prirátajú hodnoty z obdobia od poslednej aktualizácie. Vektory značiek a LDA tém sa následne nanovo normalizujú a z TF matice sa nanovo predpočíta matica TF-IDF.

Použitie podprofilov používateľov v sekciách informačného bulletinu

Na základe sekcie informačného bulletinu, pre ktorú sa vytvára zoznam odporúčaní sa využíva buď záujmový alebo expertízny podprofil používateľa, alebo oba, ako popisuje tabuľka 6.3.2.

Tabuľka 6.1: Sekcie informačného bulletinu a k nim prislúchajúce podprofily používateľa

Sekcia	Záujmový podprofil	Expertízny podprofil
Najnovšie otázky	×	–
Užitočné otázky	×	–
Otázky čakajúce na odpoveď	–	×
Populárne nezodpovedané otázky	–	×
Vysoko diskutované otázky	×	×
Vysoko diskutované odpovede	×	×
Zaujímavé odpovede	×	×

Komunitný profil používateľov

V prípade, že používateľský profil neobsahuje dostatočné množstvo aktivity pre zostavenie relevantných odporúčaní, sa použije tzv. *Komunitný profil používateľov*. Ten je presne analogický

so štandardným profilom používateľa, no skladá sa nie z aktivity konkrétneho používateľa, ale zo všetkých položených otázok a odpovedí v celom CQA systéme. Rovnako ako štandardné používateľské profily sa aj komunitný profil aktualizuje denne a s aplikáciou faktoru úpadku.

6.3.3 Zostavenie zoznamu odporúčaní

Modul pre personalizované odporúčanie podporuje dve metódy pre zostavenie zoznamu odporúčaní – štandardné personalizované odporúčanie a personalizované odporúčanie s diverzifikačnou metódou tématického vzorkovania. Tieto dve metódy sme využili pri overení našich hypotéz prostredníctvom A/B testu (viď kapitola 7).

Personalizované odporúčanie

Pri personalizovanom odporúčaní n položiek je proces pomerne priamočiary: na základe sekcie informačného bulletinu sa vyberie príslušný podprofil používateľa, z ktorého sa následne vyberie n najrelevantnejších značiek a $\left\lceil \frac{n}{2} \right\rceil$ najrelevantnejších LDA tém. Potom sa z databázy vyberú všetky otázky, položené od odoslania predošlého informačného bulletinu, ktoré patria do týchto LDA tém alebo značiek.

Prostredníctvom skalárneho súčinu nad TF-IDF maticou vybraných otázok a maticou z profilu používateľa sa potom vytvorí usporiadaný zoznam otázok, z ktorých je prvých n prezentovaných používateľovi.

Personalizované odporúčanie s diverzifikačnou metódou tématického vzorkovania

Podobne ako pri štandardnom personalizovanom odporúčaní sa najprv vyberie určitá množina značiek a LDA tém z príslušného používateľského podprofilu. Pre každú z nich sa následne rovnakou metódou zostaví zoznam odporúčaní. Tieto jednotlivé zoznamy sa potom spoja do jedného výsledného zoznamu n odporúčaní. Podrobnosti metódy tématického vzorkovania sú uvedené v kapitole 5.2.

V prípade, že metóda odporúčania s diverzifikáciou nevráti dostatočne početný zoznam odporúčaní, doplní sa zoznam o položky získané z metódy personalizovaného odporúčania. Tento prípad môže nastať hlavne v prípade denného informačného bulletinu, ak sa vyberú značky s veľmi malým počtom nových otázok.

Obe metódy tiež zabezpečujú, aby používateľovi neboli v jednom informačnom bulletine prezentované tie isté otázky viackrát.

6.4 Charakteristika dát

Našu prácu sme sa rozhodli realizovať nad dátami z komunity *Stack Overflow*, nakoľko je táto komunita zameraná na doménu, v ktorej máme hlbšie vedomosti a teda vieme lepšie posúdiť správnosť odporúčania v tejto doméne. Navrhnuté riešenie však nie je špecifické pre túto komunitu a je možné ho nasadiť v rámci celej platformy Stack Exchange, prípadne na iných CQA systémoch s podobnou štruktúrou.

Komunita Stack Overflow, ktorej dáta sme použili, je jednoznačne najväčšou a najaktívnejšou zo všetkých komunít platformy Stack Exchange a má nasledovný rozsah (*údaje sú zaokrúhlené*):

- Otázky – 16 miliónov
- Odpovede – 24 miliónov
- Používatelia – 9 miliónov
- Komentáre – 66 miliónov
- Značky – 52 tisíc
- Priemerný denný prírastok otázok – 7 tisíc
- Priemerný denný prírastok odpovedí – 7.5 tisíc
- Priemerný denný prírastok komentárov – 26 tisíc

Archívne dáta

Platforma Stack Exchange zverejňuje archív všetkých komunít vo forme XML výstupov. Dáta v tomto archíve⁴ sú pravidelne aktualizované a obsahujú kompletne používateľmi vytvorené anonymizované dáta zo všetkých komunít platformy. Všetky tieto dáta sú verejne dostupné pod licenciou *Creative Commons Attribution-ShareAlike 3.0 Unported*.

Tieto dáta sme využili v prvotnej fáze na vytvorenie základných modelov. Následne sa využívali dáta získané prostredníctvom verejného API poskytovaného platformou Stack Exchange.

Štruktúra dát v archívoch je nasledovná:

- *Badges.xml* – Obsahuje ID používateľov, názvy odznakov a čas, kedy používateľ odznak získal.
- *Comments.xml* – Obsahuje všetky komentáre spolu s informáciou o ich autoroch a príspevkoch, ku ktorým sa viažu.
- *Posts.xml* – Obsahuje informácie o všetkých príspevkoch (otázkach a odpovediach) a k nim prislúchajúce značky, ako aj aktuálne znenie príspevku

⁴<http://archive.org/details/stackexchange>

- *PostHistory.xml* – Obsahuje históriu zmien jednotlivých príspevkov, ako napr. zmenu názvu, štítkov, označenie otázky za zodpovedanú a pod.
- *Users.xml* – Obsahuje verejné údaje všetkých používateľov, ako sú meno, reputácia, webová stránka, počet hlasov a iné.
- *Votes.xml* – Obsahuje anonymizované informácie o hlasoch príspevkov.

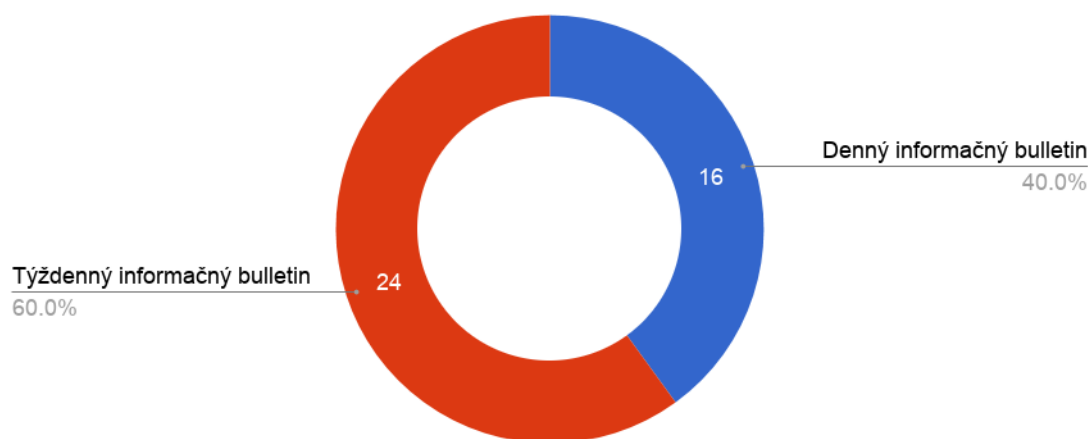
Dátový archív komunity Stack Overflow poskytuje všetok obsah od vzniku systému (cca 200GB dát vo formáte XML). Nakoľko však pre naše účely (zostavenie slovníka výrazov, natrénovanie LDA tém) nie je potrebný celý archív, rozhodli sme sa použiť len obmedzené množstvo archívnych dát, konkrétne len príspevky od 1.6.2017. Okrem nich boli tiež stiahnuté všetky príspevky, s ktorými v minulosti prišli do kontaktu odoberatelia nášho informačného bulletinu. Keďže informačné bulletiny majú charakter periodického média s veľkým dôrazom na aktuálnosť obsahu, toto obmedzenie sa len na určitú podmnožinu dát nemá na výsledky žiaden signifikantný vplyv.

7 Experimentálne overenie

Navrhnutú metódu vytvárania personalizovaných informačných bulletinov sme sa rozhodli overiť prostredníctvom nekontrolovaného online experimentu. Väčšina prác venujúcich sa odporúčaniu v doméne CQA systémov overuje svoje hypotézy v offline experimentoch na kontrolovaných vzorkách dát. Pre účely realistického overenia našej metódy, hlavne z pohľadu faktorov špecifických pre informačné bulletiny, však overovanie v offline experimente nie je dostačujúce. Online nekontrolované experimenty samozrejme vyžadujú oveľa viac námahy a ich výsledky nie sú garantované, no v prípade, že je takýto experiment úspešný, má oveľa väčší potenciál priniesť relevantné výsledky.

Experiment sme vykonávali na používateľoch z komunity Stack Overflow platformy Stack Exchange. Komunitu Stack Overflow sme pre náš experiment zvolili z viacerých dôvodov, hlavne však preto, že sa jedná o jednoznačne najaktívnejšiu komunitu z celej platformy, v ktorej bol najväčší potenciál získať čo najviac používateľov do nášho experimentu. Okrem toho zohrával úlohu v rozhodovaní aj fakt, že táto komunita sa venuje doménovej oblasti, v ktorej máme určité vedomosti, čo nám umožnilo jednoduchšie odhadnúť úroveň relevancie odporúčaného obsahu.

Do experimentu sa mohli prihlásiť všetci používatelia, ktorí majú konto na portáli Stack Overflow, prostredníctvom registračného formuláru systému StackLetter. Celkovo sa nám podarilo získať pomerne diverznú vzorku používateľov, medzi ktorými boli prítomní okrem iných aj úplni nováčikovia bez akejkoľvek predošlej aktivity, ale aj mimoriadne aktívni a dlhodobo rešpektovaní členovia komunity. Používatelia mali pri registrácii možnosť vybrať si buď denný alebo týždenný informačný bulletin. Túto voľbu tiež mohli kedykoľvek počas experimentu zmeniť.



Obr. 7.1: Podiel a počet odoberateľov variant informačného bulletinu.

System StackLetter sme propagovali všetkými nami dostupnými spôsobmi: prostredníctvom portálu StackApps¹, ktorý slúži na prezentáciu aplikácií využívajúcich API platformy Stack Exchange, taktiež prostredníctvom komunitných portálov Reddit² a Hacker News³, osobne na fakulte a tiež vo výskumnej skupine *PeWe datalys*⁴.

7.1 Metodológia experimentu

Experiment prebiehal formou A/B testovania. Na rozdiel od štandardného A/B testovania, v ktorom sa používatelia rozdelia do dvoch alebo viacerých skupín, sme náš test vykonávali na báze striedania použitých metód v časových intervaloch:

1. Kontrolná skupina

V prvej fáze experimentu, od 9. novembra 2017 do 11. marca 2018, sme používateľom generovali informačný bulletin len prostredníctvom triviálnej metódy odporúčania, ktorá vyberala otázky označené značkami, v ktorých v minulosti používateľ položil otázku, ponúkol odpoveď alebo okomentoval príspevok. Dáta z tohto časového obdobia slúžili ako kontrolná skupina pre porovnanie s ostatnými metódami.

Od 12. marca 2018 sme nasadili metódy vytvárania odporúčaní opisované v tejto práci. Metódy generovania informačných bulletinov sa striedali na týždennej báze.

2. Metóda A – Personalizované odporúčanie

Využívala sa metóda personalizovaného odporúčania opísaná v kapitole 6.3.3, pričom sa aplikovala metóda pre zabezpečenie aktuálnosti odporúčaného obsahu. V tomto prípade sa nevykonávala žiadna diverzifikácia odporúčaní.

3. Metóda B – Personalizované odporúčanie s diverzifikáciou

V tomto prípade sa zostavovali personalizované informačné bulletiny s využitím diverzifikačnej metódy tématického vzorkovania, ako aj metódy pre zabezpečenie aktuálnosti odporúčaného obsahu.

A/B testovanie formou striedania metód nám umožnilo vyhnúť sa problému s nerovnomerne rozloženými používateľmi v jednotlivých skupinách z pohľadu ich aktivity, a tiež nespôsobilo problém pri postupne sa vyvíjajúcom počte používateľov. Naopak ako potenciálny problém,

¹<https://stackapps.com>

²<https://www.reddit.com/r/stackoverflow/>

³<https://news.ycombinator.com/>

⁴<https://www.pewe.sk/datalys>

s ktorým treba pri takomto riešení počítať, sa ukázalo prirodzené postupné znižovanie záujmu a aktivity používateľov počas trvania experimentu.

7.2 Metriky hodnotenia výsledkov

Pre overovanie výsledkov experimentov sme využili nasledovné 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 odporúčaní je pre používateľa relevantná.

$$\text{Precision@N} = \frac{|\{\text{relevantne otázky v top-N}\} \cap \{\text{top-N odporucených otázok}\}|}{|\{\text{top-N odporucených otázok}\}|}$$

DCG

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 [32].

DCG položky na pozícii p v zozname odporúčaní je definované ako:

$$\text{DCG}_p = \sum_{i=1}^p \frac{2^{rel_i} - 1}{\log_2(i + 1)}$$

rel_i – relevancia i -tej položky v zozname odporúčaní.

CTR a používateľská aktivita

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. V našom experimente sme za úspešné považovali kliknutia na samotné položky alebo poskytnutie pozitívnej spätnej väzby na položku.

Okrem miery úspešných preklikov sme využili aj odvodenú metriku používateľskej aktivity, ktorá vyjadruje celkový počet kliknutí na akékoľvek odkazy v informačnom bulletine.

Predpoklady výsledkov metrík

Na základe našich hypotéz sme predpokladali nárast vo všetkých sledovaných metrikách pri porovnávaní metódy personalizovaného odporúčania s kontrolnou skupinou. Pri porovnávaní diverzifikovaného odporúčania a personalizovaného odporúčania sme predpokladali čiastočný pokles metrík relevancie odporúčaní (Precision@N, DCG), no nárast metrík používateľskej aktivity – CTR, impresie, konverzie.

7.3 Vyhodnotenie výsledkov experimentu

V realizovanom online nekontrolovanom experimente, ktorý prebiehal od 9. novembra 2017 a v ktorom boli nasadené navrhnuté metódy od 12. marca 2018, sa celkovo generovali informačné bulletiny pre 40 používateľov. Počas trvania experimentu sa z odoberania informačného bulletinu odhlásili traja používatelia. Podrobnejšie štatistiky experimentu uvádza tabuľka 7.1.

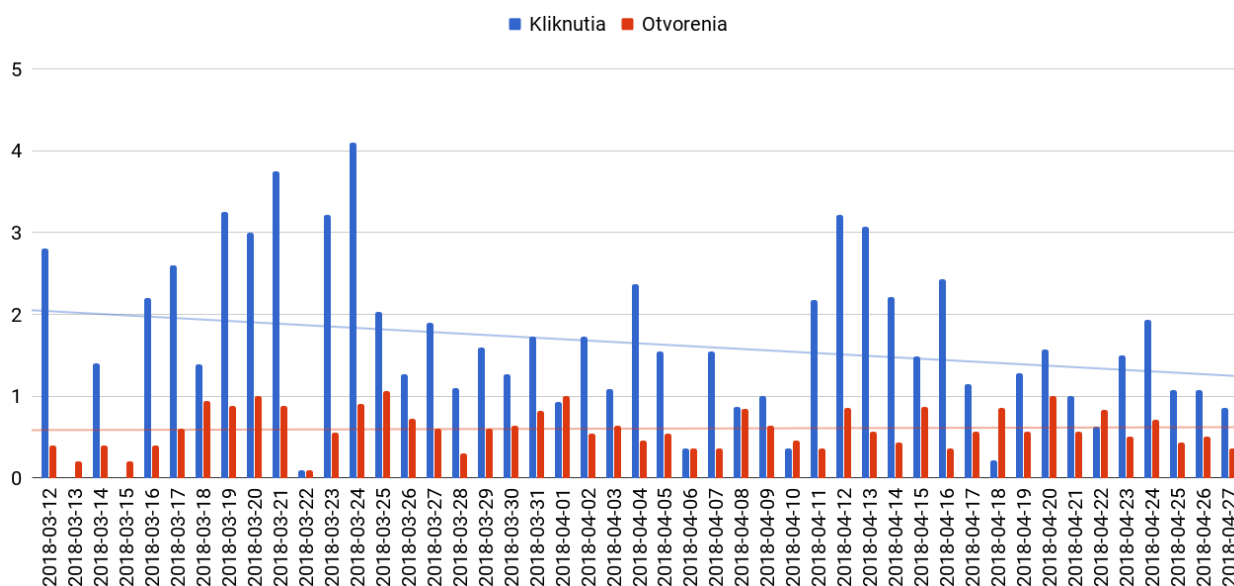
Tabuľka 7.1: Štatistika používateľskej aktivity v online experimente.

Štatistika	Počet
Odoberatelia	40
Odoslané informačné bulletiny	2400
Otvorené informačné bulletiny	750 ⁵
Celkový počet otvorení bulletinov	1050
Implicitná spätná väzba	530
Explicitná spätná väzba	960
Explicitná spätná väzba	960
Odhlásenia informačného bulletinu	3

7.3.1 Aktivita používateľov

Aktivita používateľov v informačnom bulletine sa ukázala ako problémový faktor nášho experimentu. Veľká časť používateľov bola totiž aktívna iba počas prvých dní alebo týždňov od registrácie, a ich aktivita postupom času klesala, ako to naznačuje aj trendová čiara v grafe 7.2. Zvolená metodológia striedania metód po týždňoch však tento problém do značnej miery eliminovala, vďaka čomu sa nám podarilo získať dostatočné množstvo dát o aktivite používateľov na to, aby sme vyhodnotili úspešnosť experimentu (viď tabuľka 7.1).

⁵Počet otvorení informačného bulletinu môže byť skreslený, nakoľko viaceré e-mailové klienty blokujú zobrazenie obrázkov použitých na vyhodnotenie otvorenia e-mailu. Jedná sa tak o spodnú hranicu reálneho počtu.



Obr. 7.2: Priemerná aktivita používateľov v jednotlivých dňoch.

7.3.2 Vyhodnotenie sledovaných metrík

Presnosť odporúčaní – Precision@N

Podľa očakávania dosiahla metóda personalizovaného odporúčania výrazne lepšie výsledky pozitívnej predikčnej hodnoty ako triviálne odporúčanie (viď tabuľka 7.2), konkrétne približne dvojnásobné. Rovnako podľa očakávania mala metóda diverzifikácie za dôsledok zníženie presnosti odporúčaní, ktoré však boli aj naďalej výrazne lepšie, ako v prípade triviálnej metódy.

Tabuľka 7.2: Priemerné hodnoty Precision@N.

	N = 1	N = 2	N = 3	N = 4	N = 5
Kontrolná skupina	0.0266	0.0235	0.0233	0.0232	0.0230
Metóda A	0.0546	0.0506	0.0485	0.0449	0.0473
Metóda B	0.0361	0.0356	0.0301	0.0292	0.0274

Užitočnosť odporúčaní – DCG

Metriku DCG sme pre triviálnu metódu odporúčania nepočítali, nakoľko táto metóda nepočítala konkrétne hodnoty relevancie jednotlivých položiek. Dopad diverzifikačnej metódy tematického vzorkovania na celkovú užitočnosť zoznamu odporúčaní sa ukázal ako zanedbateľný.

Tabuľka 7.3: Priemerné hodnoty DCG na pozícii p.

	p = 1	p = 2	p = 3	p = 4	p = 5
Metóda A	1.02	1.66	2.17	2.61	3.00
Metóda B	1.01	1.65	2.16	2.60	2.98

Spokojnosť používateľov – CTR a používateľská aktivita

Podľa našich predpokladov z hypotézy 1 sa potvrdilo, že zavedenie personalizovaného odporúčania do informačného bulletinu CQA systému viedlo k zvýšeniu miery jeho používania a aj celkovej používateľskej aktivity. Hodnoty CTR dosiahli v metóde A (personalizovaný informačný bulletin) takmer 2,5-násobok voči kontrolnej skupine, celková používateľská aktivita narástla až na takmer 3,5-násobok.

Predpoklady z hypotézy 2, ktorá hovorila o zvýšení používateľskej aktivity po aplikovaní diverzifikácie, sa v našom experimente nepotvrdili. Používatelia očakávali personalizované odporúčania, a preto akúkoľvek diverzifikáciu vnímali negatívne, čo sa prejavilo vo zvýšenom množstve záporných hodnotení položiek prostredníctvom explicitnej spätnej väzby. Napriek tomu metóda B dosiahla o 42% lepšie výsledky v metrike CTR voči kontrolnej skupine, a 2,7-násobný nárast v celkovej aktivite, spôsobený prevažne spomínanou explicitnou spätnou väzbou na položky, ktoré používatelia vnímali ako nerelevantné.

Namerané hodnoty týchto metrík sme overili prostredníctvom analýzy rozptylu (ANOVA), ktorá potvrdila, že sú štatisticky významné.

Výsledky týchto metrík ukazujú, že používatelia, ktorí boli súčasťou nášho online experimentu nevnímajú v prípade informačných bulletinov uzavretie do *filtračnej bubliny* ako problém, ale naopak ako žiadany stav.

Tabuľka 7.4: Priemerné hodnoty CTR a používateľskej aktivity.

	CTR	Používateľská aktivita
Kontrolná skupina	0.024	0.036
Metóda A	0.058	0.123
Metóda B	0.034	0.098
ANOVA		
F-value	12.69	34.14
P-value	3×10^{-6}	2×10^{-15}

7.3.3 Dotazník odoberateľov

Súčasťou vyhodnotenia bol aj dotazník pre odoberateľov informačného bulletinu. Dotazník od 21. apríla 2018 vyplnilo 27.5% odoberateľov. Celkovo boli odoberatelia na základe odpovedí z dotazníku spokojní s ponúkaným obsahom. 63% z nich si všimlo, že obsah bol personalizovaný a všetci respondenti považujú personalizáciu obsahu informačných bulletinov za užitočnú a za veľmi dobrý nápad.

Odoberatelia nevedeli subjektívne posúdiť, či bol pre nich ponúkaný obsah relevantný, väčšina

z nich zvolila prostrednú možnosť. 55% respondentov označilo subjektívne svoju aktivitu za podpriemernú, 45% z nich sa označilo za mierne aktívnejších.

Výsledky z dotazníku samozrejme môžu byť do určitej miery skreslené, keďže v nich respondenti prezentujú svoje subjektívne postoje. Zároveň predpokladáme, že skupina neaktívnych odoberateľov dotazník nevyplnila vôbec, preto ich postoje nemusia byť adekvátne zastúpené. Podrobné výsledky dotazníku sú priložené v dokumente v prílohe C.

7.3.4 Diskusia

8 Zhodnotenie

TODO

- zhodnotenie vysledkov - jedna hypoteza vysla, druha nie.
- online experiment bol unikatny, tazko realizovatelny, napriek tomu priniesol zaujimave vysledky
- to ze ludia chcu filter bubble je mozno este zaujimavejsie ako keby to nechceli
- diverzitu prirodzene skor vnimaju ako negativum, su zvyknuti ze kazdy im chce co najviac personalizovat, hlavne ak to tvrdi nadpis newslettera.
- historicka stigma okolo newsletterov, ludia ich moc nechcu odoberat, casto vnimaju ako len dalsi spam, a ak sa prihlasia tak po case velmi rychlo prestanu byt aktivni.

– Future work

- explorovat *stigmatu* okolo newsletterov, ine metody ozivenia”standardnej personalizacie, aj ked sa momentalne zda ze pouzivatelja chcu hlavne cisto personalizovane veci.
- ine vyuzitie newsletterov v CQA, aj mimo personalizacie.

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] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors. *Recommender Systems Handbook*. Springer US, 2011.
- [10] 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.
- [11] 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.
 - [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] Fei Xu, Zongcheng Ji, and Bin Wang. Dual role model for question recommendation in community question answering. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12*. ACM Press, 2012.
- [31] Matúš Salát. Segmentovanie používateľov pre personalizáciu informačných bulletinov v cqa systémoch. Diplomová práca, Fakulta informatiky a informačných technológií, STU Bratislava, 2018.

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

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 rámci 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	<ul style="list-style-type: none">– Príprava databázy a aktualizácie modulu.– Vytvorenie modelov pre získavanie spätnej väzby od používateľov.– Prvotná dátová analýza.
3-4. týždeň semestra	<ul style="list-style-type: none">– Príprava rozhrania pre odoberanie informačných bulletinov.– Generovanie informačných bulletinov na základe prvotného modelu.– Nasadenie systému a otestovanie v reálnych podmienkach.
5-8. týždeň semestra	<ul style="list-style-type: none">– 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	<ul style="list-style-type: none">– Overenie a vyhodnocovanie informačných bulletinov.– Nasadenie a porovnanie metód diverzifikácie.– Počiatočné overenie výsledkov.– V prípade neúspechu online experimentu, plánovanie offline experimentov.

Zhodnotenie

Navrhnutý plán práce v rámci diplomového projektu I sa nám do značnej miery podarilo dodržať. Príprava a implementácia infraštruktúry systému StackLetter však zabrala viac času, ako sme predpokladali, takže došlo k určitému časovému sklzu. Voči verzii z DP1 sme tiež prepracovali návrh metódy diverzifikovaného odporúčania. Samotnú implementáciu odporúčania a diverzifikácie sme posunuli do nasledujúceho letného semestra. Celkovo sme s aktuálnym stavom spokojní, napriek tomu, že sme nedodrжали predpokladaný plán, keďže sa implementácia infraštruktúry a kooperácia s externými systémami Stack Exchange ukázala náročnejšia, ako boli naše predpoklady.

A.3 Plán práce na diplomovom projekte III

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

1-2. týždeň semestra	– Pokračovanie online nekontrolovaného experimentu. – Spracovanie dát, vytvorenie modelov používateľov a otázok.
3-4. týždeň semestra	– Príprava metódy proporčnej diverzifikácie odporúčaní – Vyhodnotenie úspechu/neúspechu online experimentu
5-6. týždeň semestra	– Príprava metódy diverzifikácie prostredníctvom tématického vzorkovania – V prípade neúspechu online experimentu, príprava kontrolovaného experimentu
7-8. týždeň semestra	– Vyhodnocovanie výsledkov experimentov – Písanie diplomovej práce
9-10. týždeň semestra	– Písanie diplomovej práce

Zhodnotenie

Navrhnutý plán práce v rámci diplomového projektu III sa nám podarilo dodržať. Online experiment bol úspešný, preto nebolo potrebné pripravovať kontrolovaný experiment.

B Technická dokumentácia systému

B.1 Modul pre registráciu nových používateľov

Tento modul predstavuje webovú aplikáciu určenú na registráciu nových používateľov systému StackLetter. Implementovaná je v jazyku PHP 7.1 s použitím MVC/MVP webového frameworku Nette 2.4. Ďalej popisujeme niektoré dôležité komponenty tohto modulu.

HomepagePresenter

Táto trieda zabezpečuje hlavnú funkcionálnosť stránky – autentifikáciu a registráciu používateľov, ako aj manažment používateľského konta. Autentifikácia používateľa sa vykonáva prostredníctvom Stack Exchange API s využitím protokolu OAuth 2.0. Zabezpečuje tiež funkcionálnosť kontaktného formuláru.

SubscriptionPresenter

Spolu s príslušnou modelovou triedou *SubscriptionModel* implementuje možnosť odhlásenia sa z odoberania informačného bulletinu, generovanie *unsubscribe* odkazov pre jednotlivé informačné bulletiny, zbieranie spätnej väzby po odhlásení sa z odoberania a možnosť opätovného prihlásenia sa.

AsyncJobProcessor

Táto trieda zabezpečuje asynchrónne spracovanie dlhotrvajúcich úloh. Vďaka asynchrónnemu spracovaniu týchto úloh ostáva odozva používateľského rozhrania rýchla. Konkrétne vykonáva odoslanie uvítacieho e-mailu používateľovi po úspešnej registrácii prostredníctvom služby SendGrid s využitím protokolu SMTP, a následné stiahnutie používateľského profilu prostredníctvom Stack Exchange API. Samotná webová aplikácia zadáva tomuto komponentu úlohy na asynchrónne spracovanie prostredníctvom Redis fronty.

B.2 Modul pre generovanie a odosielanie informačných bulletinov

Tento modul je zodpovedný za samotné zostavovanie a odosielanie informačných bulletinov jednotlivým zaregistrovaným používateľom. Skladá sa z dvoch samostatných častí – komponentu na zostavovanie informačných bulletinov a komponentu na odosielanie informačného bulletinu.

Oba komponenty medzi sebou asynchrónne komunikujú prostredníctvom Redis fronty.

Modul je navrhnutý genericky, aby neoblo potrebné do neho zasahovať v prípade pridania nových sekcií informačných bulletinov, prípadne nových typov obsahu. Pre každý typ obsahu je potrebné iba definovať HTML šablónu.

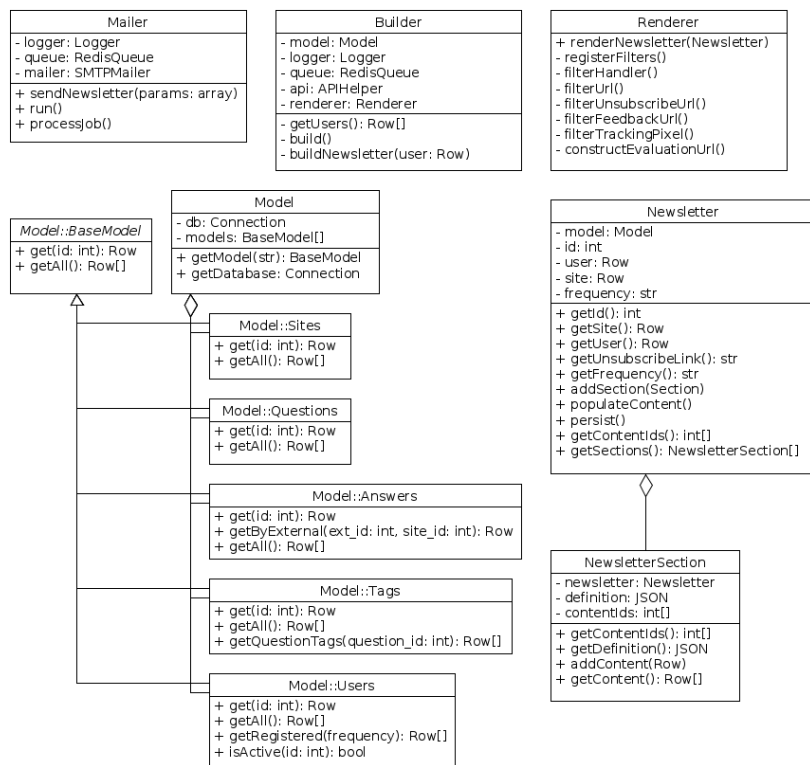
Všetky hypertextové odkazy, ktoré sa nachádzajú v informačnom bulletine, smerujú na evaluačný modul, ktorý prostredníctvom kliknutí na odkazy zaznamenáva používateľovu implicitnú a explicitnú spätnú väzbu a následne je používateľ presmerovaný na reálny cieľ odkazu. Každý informačný bulletin tiež obsahuje tzv. sledovací pixel (angl. *tracking pixel*), čo je transparentný obrázok veľkosti 1x1 pixelov, načítavaný externe z evaluačného modulu, ktorý nám umožňuje detekovať otvorenie informačného bulletinu používateľom.

Builder – komponent pre zostavovanie informačného bulletinu

1. Prostredníctvom *cronu* sa tento komponent spúšťa každý deň pre vygenerovanie denných informačných bulletinov, ako aj raz za týždeň pre generovanie týždenných bulletinov.
2. Na základe zvolenej frekvencie sa vyberie zoznam používateľov, pre ktorých sa budú generovať bulletin.
3. Prostredníctvom REST API si komponent vyžiada štruktúru bulletinu pre daného používateľa. Odpoveď obsahuje zoznam sekcií, ich nadpis a popis, ako aj URL adresu REST API koncového bodu, ktorý má pre danú sekciu vrátiť obsah.
4. Prostredníctvom REST API si komponent vyžiada zoznam otázok (alebo iného obsahu) pre jednotlivé sekcie informačného bulletinu.
5. Následne komponent zostaví kompletnú štruktúru celého informačného bulletinu so všetkými sekciami a ich obsahom, a túto štruktúru odovzdá šablónovaciemu komponentu, ktorý zostaví výsledný HTML kód informačného bulletinu.
6. Informácia o vygenerovaní nového informačného bulletinu sa vloží do Redis fronty, kde čaká na odoslanie.

Mailer – komponent pre odosielanie informačného bulletinu

Tento komponent postupne vyberá z Redis fronty zoznam vygenerovaných informačných bulletinov, a využívajúc SMTP API služby SendGrid odosiela používateľom jednotlivé informačné bulletin, na ktorých odoberanie sa prihlásili. Tento komponent je spustený na serveri ako služba, a je k dispozícii neustále.



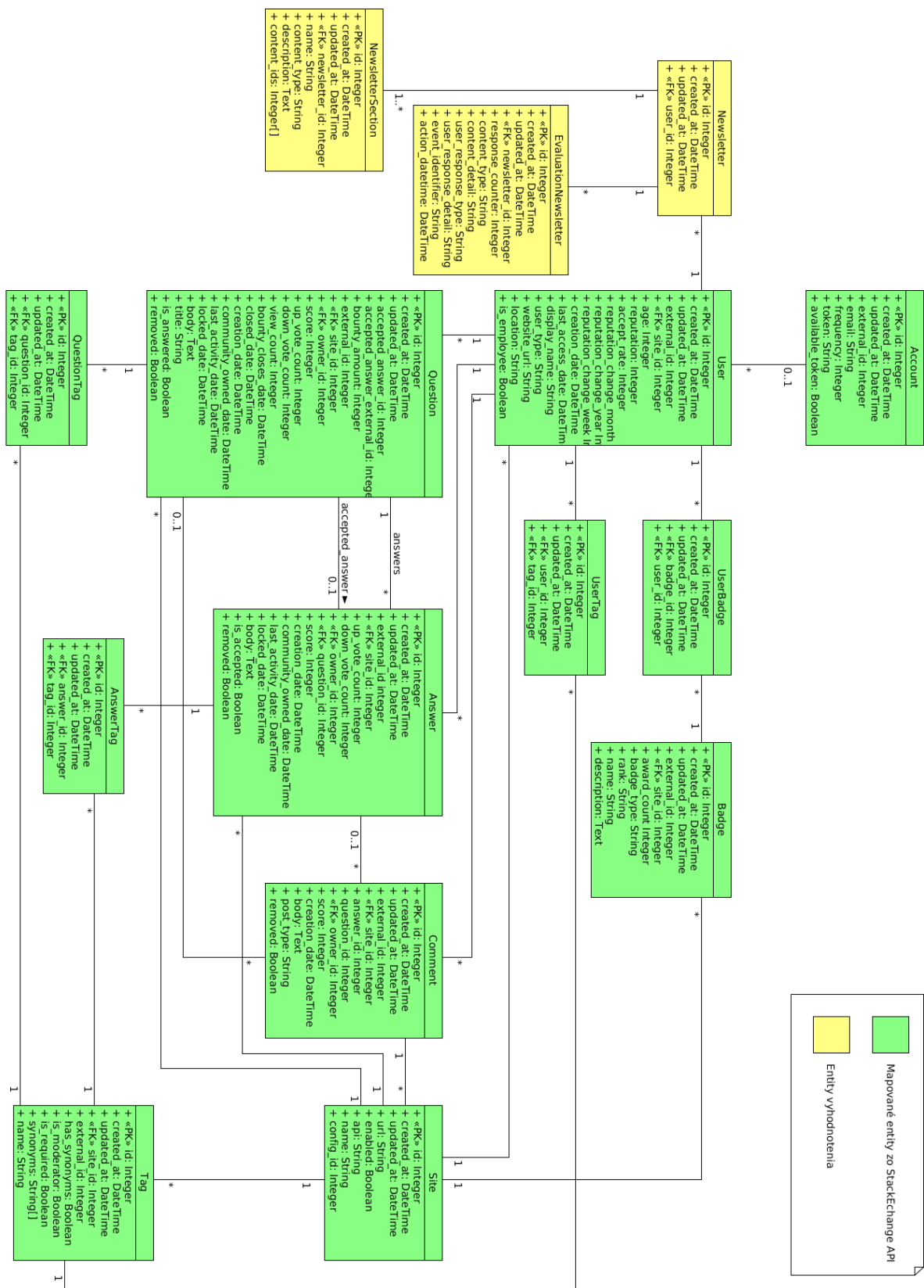
Obr. B.1: Schéma modulu pre generovanie a odosielanie informačných bulletinov

B.3 Modul pre spracovanie statickej zálohy databázy

Tento modul číta *XML dump* Stack Exchange databázy a generuje príkazy SQL dialektu PostgreSQL pre uloženie dát do databázového modelu systému StackLetter (príloha B.4). Zároveň tiež vykonáva konverziu mapovania jednotlivých entít zo Stack Exchange reprezentácie.

Import dát je vykonávaný dávkovo a v poradí podľa závislostí jednotlivých entít – najprv sa importujú používatelia, následne otázky, odpovede, komentáre, značky a nakoniec odznaky.

B.4 Databázový model

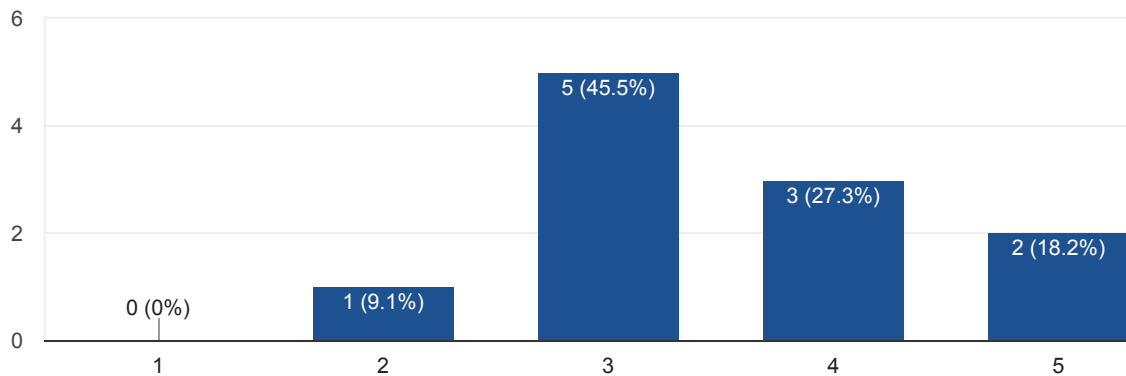


Obr. B.2: Databázový model systému StackLetter.

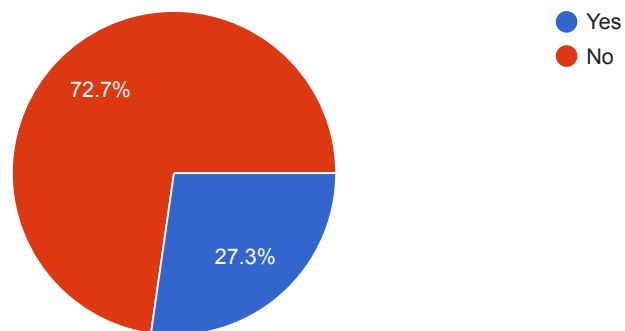
C Výsledky dotazníku odoberateľov

Subscriber feedback survey

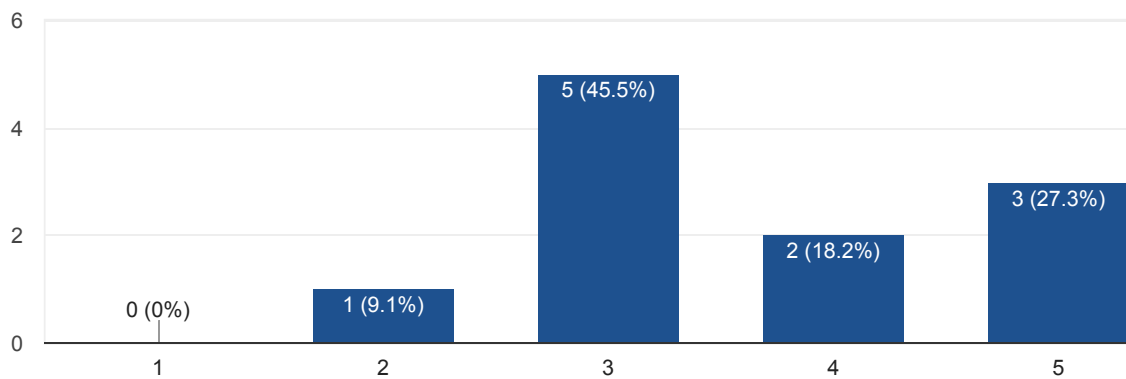
How did you like the sections provided in your newsletters?



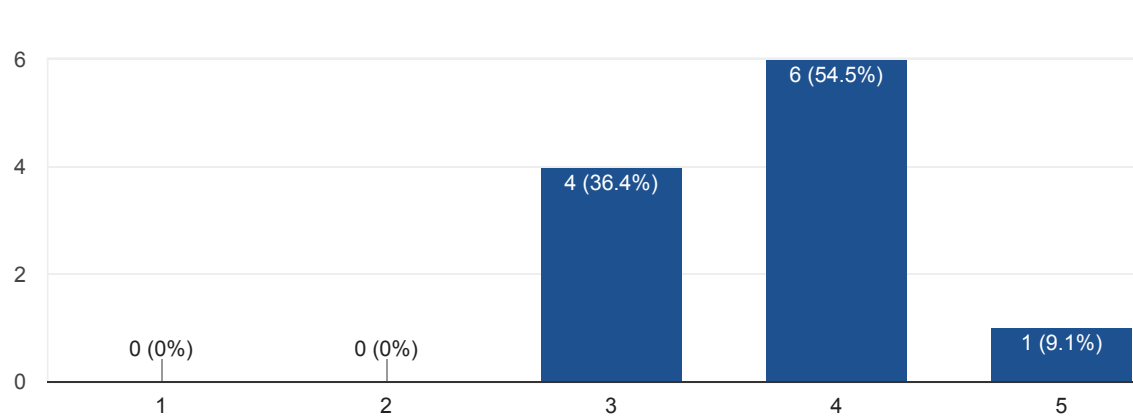
Did you notice the adaptive structure of the newsletter?



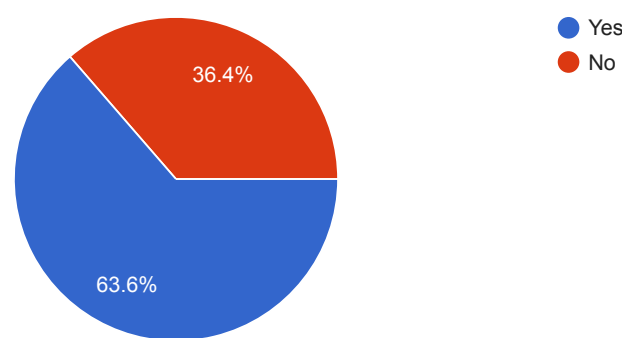
Do you consider adaptive structure to be useful?



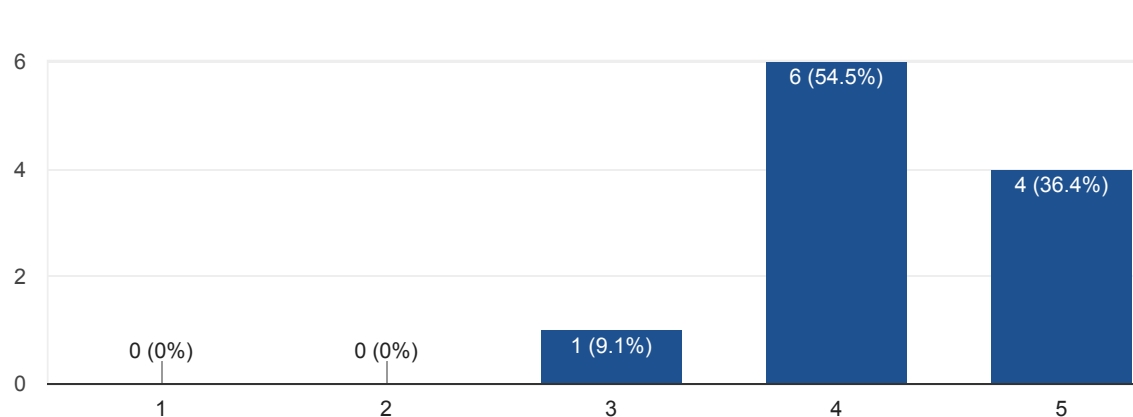
How did you like the content provided in your newsletters?



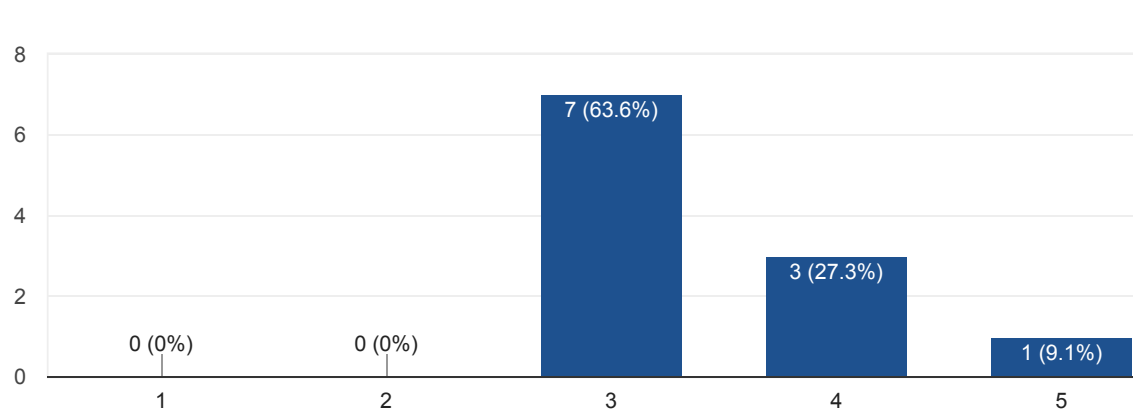
Did you notice the personalization of the newsletter content?



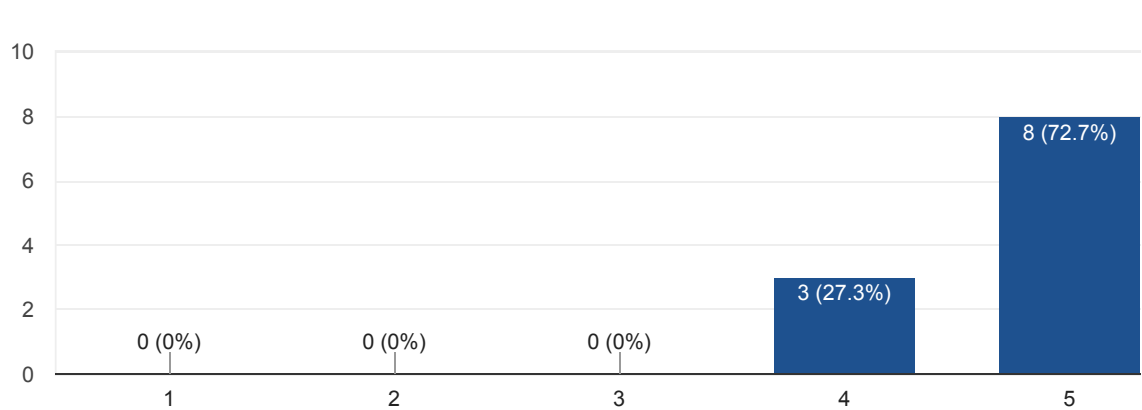
Did you find the content personalization useful?



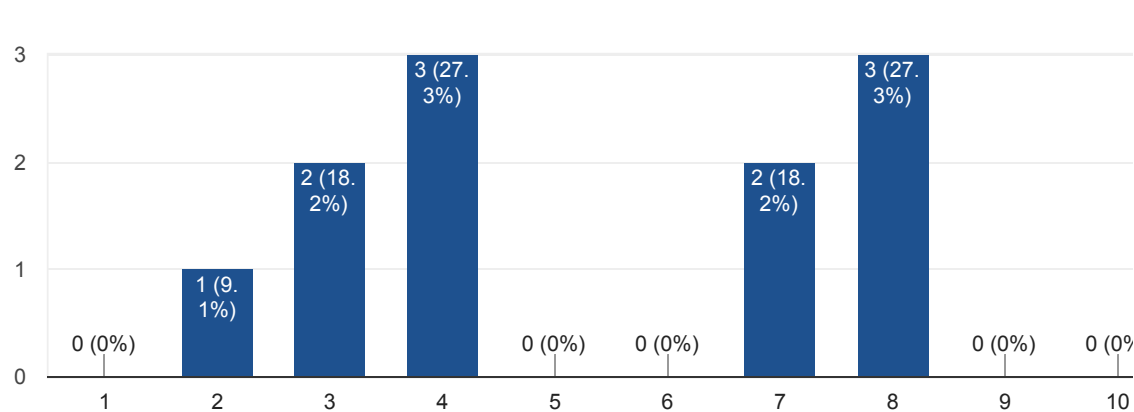
Do you think the personalized content was relevant to you?



Do you think a personalized newsletter is a good idea?



Please evaluate your activity within the newsletter



D Príspevok na konferencii IIT.SRC 2018

Improving Diversity and Freshness of Newsletters in Community Question Answering Systems

Martin ŠRANK*

*Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
xsrankm@stuba.sk*

Abstract. Newsletters represent a standard way to inform users in online communities about new or interesting content. Their importance is even greater in communities producing large amounts of user-created data, such as community question answering (CQA) systems. Nevertheless, many popular CQA systems only offer generic newsletters. 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 focus on improving the diversity and freshness of recommendations as a way to prevent filter bubbles and improve user satisfaction and engagement.

1 Introduction

Newsletters, even today, are one of the prevalent ways of distributing news and content to users of online communities. They are mostly used to promote new content, discounted products or special offers, but also as a way to motivate the user to visit the site again. There are couple of methods being used for building the newsletters - manual content selection by an editor, automatic generation of generic newsletters or automatically generated personalized newsletters.

Newsletters are even more important in online communities producing large quantities of user-generated content, a prime example of which are Community Question Answering Systems, or CQAs. We argue that automated personalized newsletters using content recommendation have the most potential to improve user satisfaction in these kinds of communities.

However, many popular CQAs currently only offer generic newsletters, which may not be as engaging for their users, or no newsletters at all. For example, Yahoo! Answers¹ does not offer any kind of newsletter, and individual sites on the Stack Exchange platform² only offer generic newsletters with no content personalized for a specific user. Only Quora³ offers a personalized newsletter to its users, but no details about it are known.

Personalization and recommendation in CQAs is a widely researched topic [5]. In spite of that, we are not aware of any existing work aimed at personalization or recommendation utilized in sending newsletters. Moreover, we argue that only personalization is not enough. Therefore the focus of our work is mainly on improving the diversity and freshness of the recommended content. We approach diversification as an effective way to prevent the standard problem of filter bubbles which is present in most recommendation systems and also as a way to improve user engagement and satisfaction.

We propose a method of automated generation of personalized newsletters for CQA systems with focus on diversity and freshness of the recommended content. We are using question recommendation for generating personalized content and a method of thematic sampling to achieve content diversity. We also consider user interest and user expertise separately as opposed to other works. Our method is designed for use on the Stack Exchange platform, but can be easily adapted to any other CQA system.

* Master study programme in field: Information Systems

Supervisor: Dr. Ivan Srba, Institute of Informatics, Information Systems and Software Engineering, Faculty of Informatics and Information Technologies STU in Bratislava

¹ <https://answers.yahoo.com>

² <https://stackexchange.com>

³ <https://www.quora.com>

2 Related Work

Current research in the domain of CQA systems [5] focuses mainly on user behaviour, question routing and recommendation and quality of questions and answers in these systems. However, we are not aware of any research into the role of newsletters in CQA systems.

Question recommendation and routing

There are two ways of recommending questions to the users of a CQA system. The more traditional *pull* method, used in question recommendation, presents a list of recommended relevant questions to the user based on their implicit or explicit request. An opposite method, *push*, is used in question routing, where the process starts with a model of an unanswered question and the system tries to route the question to a specific user who has the most potential to answer it correctly.

Most of the research into content recommendation in CQA systems is focused on routing questions to experts [2]. However, question routing creates a problem of overloading the experts. Thus authors in [3, 4] explore the idea of routing new questions to larger groups of users instead. In our work, we instead use the more traditional method of question recommendation, since it is not as well researched in the area of CQA systems.

Diversification

Thematic diversification is a method which helps to better balance the entire spectrum of users' interests or expertise. Although it can have a negative effect on the accuracy of recommendations, this method achieves better perceived user satisfaction [7].

Authors in [8] propose a method of diversification based on selecting lists with low intra-list similarity. Another method presented in [1] uses diversification based on proportionality.

In the area of CQA systems, the most similar work to ours is presented in [6], where authors explore diversification based on a similar method of thematic sampling, which uses LDA, lexical and categorical models of questions to construct a probability tree which represents the user profile. In contrast, we adopt a different approach to constructing the final recommendation lists and we don't use probabilistic distributions in the user profile.

To summarize, there is very little research in the domain of question recommendation and freshness in CQA systems, as well as no research with regards to the use of newsletters. This is in contrast to the potential of newsletters and their current state in CQA systems, particularly in Stack Exchange.

3 Proposed Method

We propose a method for assembling personalized newsletters for users of a CQA system, which uses

content-based filtering and question recommendation to provide content recommendations to users.

Key features of the method are as follows:

- We use the method of content-based filtering, as it is more effective than collaborative filtering for users with low activity. We start with a user model, which is assembled from the models of questions they interacted with.
- Diversification is performed as a thematic prefilter before the recommendation process - a separate list of recommendations is created for each individual topic.
- The final list of recommended content is created by merging of the partial lists from previous step.
- When recommending, we consider freshness, as well as diversity of the content.

A high-level overview of the method process is shown in Figure 1.

Question model

The question model consists of three independent models, each considering the question from a different perspective:

1) **Tag-based model** represents the question on the highest level as belonging to particular tags in the CQA system. Each question can belong to multiple tags, represented as an N -dimensional vector of values 0 or 1.

2) **Thematic model** of the question uses the method of Latent Dirichlet Allocation (LDA) to assign each question to multiple latent topics. To optimize the model we ignore topics which fall in the lower 25% of overall distribution of topics for a given question.

3) **Lexical model** of the question uses TF-IDF vector representing the distribution of individual terms in the question title and body as bag-of-words model. To preprocess the texts we utilize lemmatization and stop-word removal in both thematic and lexical models.

User model

The user model is based on the question model and is assembled from the models of questions, with which the user interacted in the system. Unlike most of the previous works on recommendations in CQA systems, we distinguish between user's expertise and interests, and model these factors individually in two separate models with the same structure.

The interest model is constructed from the questions which were asked by the user, or which the user marked as their favourite. Each of these questions will contribute to the interest model with the same weight.

The expertise model is constructed from the questions which the user answered. A positive score of

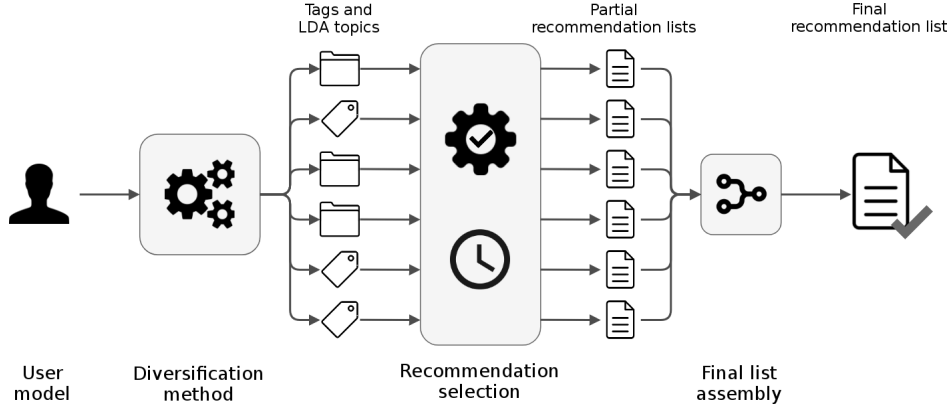


Figure 1. Personalized newsletter assembly schema

their answer will denote their expertise and a negative score will denote the lack of expertise in the question topic. Furthermore, if the user's answer is marked as accepted, it contributes to the model with a weight coefficient of 1.75, determined from the ratio of total answers for a question with an accepted answer.

All the questions to which the user added a comment also contribute to their model. However, since it is not possible to determine interest or expertise from an act of commenting, such questions contribute to both models with a weight coefficient of 0.3, since questions and answers have on average three comments.

The separation of user's interests and expertise allows us to provide a better level of personalization, as we can more accurately provide them with questions they would find interesting, as well as questions for which they have the potential to provide a good answer.

Content recommendation and freshness

The process of recommending questions to a user will use a dot product to match a user model with models of questions.

To represent the gradual change of users' interest over time, we introduce an exponential decaying factor to the process of creating the user model. On each user model update, we decrease the weight of existing data in the model proportionally to the amount of activity since the last update.

Thematic sampling diversification

As illustrated by Figure 2, the process of diversifying the recommended content is designed as follows:

1. For each user, we randomly select from their model k_1 tags and k_2 topics, only considering tags and topics above the median to suppress low-relevance tags and topics, where $k_1 + k_2 = n$ and

n is the number of recommendations in the final list.

2. For each tag and topic we then construct a list of n recommendations using the aforementioned method, while only considering questions from the given tag or topic.
3. Then we randomly sample items from the individual lists to the final list of recommendations, where the probability of selecting a question from a given list is proportional to the relevance of the particular tag or topic to the user.
4. Specific items from the individual lists will be randomly selected from *top-M* questions, where M is the relative relevance of the topic or tag.

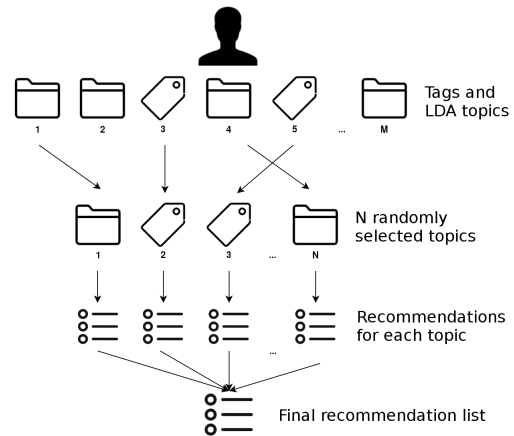


Figure 2. Schema of the recommendation diversification method of thematic sampling

4 Evaluation

We evaluate our proposed method of personalized content recommendation and diversification in an online uncontrolled experiment on the users of the Stack Overflow site from the Stack Exchange Network. This experiment has the form of a regular newsletter that every user of the platform can subscribe to.

The effectiveness of the proposed method is currently evaluated in a time-based A/B test:

1. Control experiment - Users received a newsletter with simple personalization based on tags.
2. Method A - Users receive a personalized newsletter without content diversification.
3. Method B - Users receive a personalized newsletter using the thematic sampling diversification method.

As of this writing, we are in the middle of the experiment and we alternate between methods A and B on a weekly basis, with the control experiment as a baseline, which was already concluded. We currently have more than 30 individual users subscribed to our newsletter, with more than 1800 sent daily and weekly newsletters. The subscribers have since provided more than 1300 instances of feedback.

Online experiments such as this are very rare in the domain of CQA systems, as most of the existing research was evaluated on offline data. For the purposes of this experiment we implemented an experimental infrastructure named StackLetter which was deployed in early November of 2017.

Our hypothesis states that by recommending questions from a wider area of users' interests and by considering their freshness we will avoid the problem of filter bubble, thus achieving higher degree of user's interest and activity.

Metrics

To evaluate the proposed method we compare its results with the non-diversified personalization method, as well as the baseline from the control experiment. We use the following metrics to evaluate different aspects of our method:

- Click-through Rate - To measure the effectiveness of the method and user engagement.
- Discounted Cumulative Gain - To measure the quality of the ranking.
- Precision@N - Precision of the recommendation on position N .
- Intra-list similarity - To measure the level of diversity in the recommendation lists

We expect to see a slight drop in precision and

DCG of the recommendations when using our diversification method, but we expect CTR to be higher, thus proving higher user engagement.

5 Conclusion

In this work, we proposed a method for automatic creation of personalized newsletters for CQA systems with a focus on improving the diversity and freshness of the recommendations. To the best of our knowledge, this is the only research into the use of newsletters in CQA systems. We also use question recommendation as opposed to the more widely researched method of question routing and we consider users' expertise and interests separately.

Finally, we evaluate our methods in an online uncontrolled experiment on real users, using our experimental infrastructure StackLetter, which is designed universally to support any future research in the domain of newsletters in CQA systems.

References

- [1] Dang, V., Croft, W.B.: Diversity by proportionality. In: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR 12*, ACM Press, 2012.
- [2] Li, B., King, I., Lyu, M.R.: Question routing in community question answering. In: *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM 11*, ACM Press, 2011.
- [3] Pal, A.: Metrics and Algorithms for Routing Questions to User Communities. *ACM Transactions on Information Systems*, 2015, vol. 33, no. 3, pp. 1–29.
- [4] Pal, A., Wang, F., Zhou, M.X., Nichols, J., Smith, B.A.: Question routing to user communities. In: *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM 13*, ACM Press, 2013.
- [5] Srba, I., Bielikova, M.: A Comprehensive Survey and Classification of Approaches for Community Question Answering. *ACM Transactions on the Web*, 2016, vol. 10, no. 3, pp. 1–63.
- [6] Szpektor, I., Maarek, Y., Pelleg, D.: When relevance is not enough. In: *Proceedings of the 22nd international conference on World Wide Web - WWW 13*, ACM Press, 2013.
- [7] Zhang, F.: Improving recommendation lists through neighbor diversification. In: *2009 IEEE International Conference on Intelligent Computing and Intelligent Systems*, IEEE, 2009.
- [8] Ziegler, C.N., McNee, S.M., Konstan, J.A., Lausen, G.: Improving recommendation lists through topic diversification. In: *Proceedings of the 14th international conference on World Wide Web - WWW 05*, ACM Press, 2005.

E Elektronické médium

Elektronické médium priložené k dokumentu má nasledovnú štruktúru:

/code

- Zdrojové kódy samotných implementovaných modulov

/doc

- Priebežná správa o riešení DP2 s anotáciami v slovenskom a anglickom jazyku

/doc/latex

- L^AT_EX zdrojové súbory dokumentácie

/doc/bibtex

- BibTeX súbor s použitými referenciami

readme.txt

- opis obsahu elektronického média