

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-182905-55695

Bc. Elena Štefancová

ODPORÚČANIE ZOHLADŇUJÚCE ČASOVÉ ASPEKTY POUŽÍVATEĽOV A POLOŽIEK

Diplomová práca

Študijný program: Inteligentné softvérové systémy

Študijný odbor: 9.2.5 Softvérové inžinierstvo (hlavný) a 9.2.8 Umelá inteligencia (vedľajší)

Miesto vypracovania: Ústav informatiky, informačných systémov a softvérového inžinierstva, FIIT STU v Bratislave

Vedúci práce: Ing. Ivan Srba, PhD.

máj 2019

Zadanie diplomovej práce

Meno študenta: **Bc. Elena Štefancová**

Študijný program: Inteligentné softvérové systémy

Študijný odbor: Softvérové inžinierstvo – hlavný študijný odbor
Umelá inteligencia – vedľajší študijný odbor

Názov práce: **Odporúčanie zohľadňujúce časové aspekty používateľov a položiek**

Samostatnou výskumnou a vývojovou činnosťou v rámci predmetov Diplomový projekt I, II, III vypracujte diplomovú prácu na tému, vyjadrenú vyššie uvedeným názvom tak, aby ste dosiahli tieto ciele:

Všeobecný cieľ:

Vypracovaním diplomovej práce preukážete, ako ste si osvojili metódy a postupy riešenia relatívne rozsiahlych projektov, schopnosť samostatne a tvorivo riešiť zložité úlohy aj výskumného charakteru v súlade so súčasnými metódami a postupmi študovaného odboru využívanými v príslušnej oblasti a schopnosť samostatne, tvorivo a kriticky pristupovať k analýze možných riešení a k tvorbe modelov.

Špecifický cieľ:

Vytvorte riešenie zodpovedajúce návrhu textu zadania, ktorý je prílohou tohto zadania. Návrh bližšie opisuje tému vyjadrenú názvom. Tento opis je záväzný, má však rámcový charakter, aby vznikol dostatočný priestor pre Vašu tvorivosť.

Riadte sa pokynmi Vášho vedúceho.

Pokiaľ v priebehu riešenia, opierajúc sa o hlbšie poznanie súčasného stavu v príslušnej oblasti, alebo o priebežné výsledky Vášho riešenia, alebo o iné závažné skutočnosti, dospejete spoločne s Vaším vedúcim k presvedčeniu, že niečo v texte zadania a/alebo v názve by sa malo zmeniť, navrhnete zmenu. Zmena je spravidla možná len pri dosiahnutí kontrolného bodu.

Miesto vypracovania: Ústav informatiky, informačných systémov a softvérového inžinierstva, FIIT STU v Bratislave

Vedúci práce: **Ing. Ivan Srba, PhD.**

Termíny odovzdania:

Podľa harmonogramu štúdia platného pre semester, v ktorom máte príslušný predmet (Diplomový projekt I, II, III) absolvovať podľa Vášho študijného plánu

Predmety odovzdania:

V každom predmete dokument podľa pokynov na www.fiit.stuba.sk v časti:
home > Informácie o > štúdiu > harmonogram štúdia > diplomový projekt.

V Bratislave dňa 12. 2. 2018

SLOVENSKÁ TECHNICKÁ UNIVERZITA
V BRATISLAVE
Fakulta informatiky a informačných technológií
Iľkovičova 2, 842 16 Bratislava 4

1

prof. Ing. Pavol Návrat, PhD.
riaditeľ Ústavu informatiky, informačných systémov
a softvérového inžinierstva

Návrh zadania diplomovej práce

Finálna verzia do diplomovej práce¹

Študent:

Meno, priezvisko, tituly: Elena Štefancová, Bc.
Študijný program: Inteligentné softvérové systémy
Kontakt: elenastefancova@gmail.com

Výskumník:

Meno, priezvisko, tituly: Ivan Srba, Ing. PhD.

Projekt:

Názov: Odporúčanie zohľadňujúce časové aspekty používateľov a položiek
Názov v angličtine: Recommendation Taking the Time Aspects of Users and Items into Account
Miesto vypracovania: Ústav informatiky, informačných systémov a softvérového inžinierstva, FIIT STU, Bratislava
Oblasť problematiky: Odporúčanie, analýza dát

Text návrhu zadania²

Úlohou odporúčacích systémov je na základe predchádzajúcej aktivity používateľov predikovať ich záujem o nové položky (napr. produkty v elektronickom obchode). Vo väčšine systémov sa ale berie ohľad najmä na používateľov a položky, už menší na aspekty vychádzajúce z kontextu, akými sú čas, lokalita či aktuálna spoločnosť iných ľudí. Pre odporúčania pritom predovšetkým časové aspekty môžu hrať významnú rolu. V prípade používateľa to môže byť zmena jeho preferencií. V prípade položky to môže byť jej aktuálnosť (napr. pri odporúčaní správ na spravodajskom portáli alebo pri vytváraní týždenného informačného bulletinu) alebo periodicita (napr. pri odporúčaní sezónnych položiek).

Analyzujte existujúce systémy zohľadňujúce časové aspekty používateľa a odporúčanej položky (angl. Time-aware Recommender Systems). Navrhňte vlastnú metódu na odporúčanie, ktorá zohľadní časové aspekty používateľa a odporúčaných položiek pre vygenerovanie presnejšieho odporúčania. Navrhnuté riešenie experimentálne overte vo zvolenej doméne.

¹ Vytlačiť obojstranne na jeden list papiera

² 150-200 slov (1200-1700 znakov), ktoré opisujú výskumný problém v kontexte súčasného stavu vrátane motivácie a smerov riešenia

Literatúra³

- ANELLI, V. et al.: An analysis on time- and session-aware diversification in recommender systems. In: Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17), ACM Press, 2017, s. 270-274.
- LOMMATZSCH, A. KILLE, B. ALBAYRAK, S.: Incorporating Context and Trends in News Recommender Systems. In: Proceedings of the International Conference on Web Intelligence (WI '17), ACM Press, 2017, s. 1062-1068.

Vyššie je uvedený návrh diplomového projektu, ktorý vypracoval(a) Bc. Elena Štefancová, konzultoval(a) a osvojil(a) si ho Ing. Ivan Srba, PhD. a súhlasí, že bude takýto projekt viesť v prípade, že bude pridelený tomuto študentovi.

V Bratislave dňa 30.12.2017



Podpis študenta



Podpis výskumníka

Vyjadrenie garanta predmetov Diplomový projekt I, II, III

Návrh zadania schválený: áno / nie⁴

Dňa: 12. 2. 2018



Podpis garanta predmetov

³ 2 vedecké zdroje, každý v samostatnej rubrike a s údajmi zodpovedajúcimi bibliografickým odkazom podľa normy STN ISO 690, ktoré sa viažu k téme zadania a preukazujú výskumnú povahu problému a jeho aktuálnosť (uvedte všetky potrebné údaje na identifikáciu zdroja, pričom uprednostnite vedecké príspevky v časopisoch a medzinárodných konferenciách)

⁴ Nehodiace sa prečiarknite

ČESTNÉ PREHLÁSENIE

Čestne vyhlasujem, že som túto prácu vypracovala samostatne, na základe konzultácií a s použitím uvedenej literatúry.

V Bratislave, 9.5.2019

.....

Bc. Elena Štefancová

POĎAKOVANIE

Ďakujem svojmu vedúcemu práce Ing. Ivanovi Srbovi, PhD. a ostatným členom výskumnej skupiny PeWe za cennú spätnú väzbu počas celého trvania diplomového projektu.

Bc. Elena Štefancová

Anotácia

Slovenská technická univerzita v Bratislave

FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLOGIÍ

Študijný program: Inteligentné softvérové systémy

Autor: Bc. Elena Štefancová

Diplomová práca: Odporúčanie zohľadňujúce časové aspekty používateľov a položiek

Vedúci diplomovej práce: Ing. Ivan Srba, PhD.

máj, 2019

Práca sa venuje problematike odporúčacích systémov, konkrétne tým, ktoré sa zameriavajú na zapracovanie kontextuálnych aspektov. Našou snahou je najmä zlepšiť odporúčacie systémy o rozšírenie zahŕňajúce temporálne aspekty ako jeden z dôležitých atribútov pri generovaní odporúčaní bodov záujmu v doméne na lokalite založených sociálnych sietí.

Za týmto účelom sme navrhli metódy, ktoré skúmajú dlhodobé trendy v popularite kategórií a vplyv sezónnosti na kategórie v závislosti od geografickej oblasti. Využívame pritom temporálne modelovanie pomocou matice črt položiek, temporálny po-filter a pred-filter a tiež geografické po-filtrovanie na základe lokality predchádzajúcich transakcií používateľa. Základom odporúčania je faktorizácia matíc kolaboratívnej povahy.

Výsledky na dátovej sade z Yelp-u naznačujú, že dlhodobé trendy nie sú sľubnou oblasťou pre zlepšenie úspešnosti, výsledky pre lokálnu sezónnosť sú významné. Dokazujú, že implementácia geografického i temporálneho kontextu v podobe lokálnej sezónnosti skutočne zlepšuje hodnoty presnosti a pokrytia, ak sa jedná o oblasti, kde je počasie celoročne silno ovplyvnené ročnými obdobiami, a preto je vhodné pri ich zahrnutí uvažovať lokalitu. Najlepšie výsledky boli dosiahnuté pre kontextuálne modelovanie.

Annotation

Slovak University of Technology Bratislava

FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES

Degree Course: Intelligent Software Systems

Author: Elena Štefancová

Master's Thesis: Recommendation Taking the Time Aspects of Users and Items into Account

Supervisor: Dr. Ivan Srba

2019, May

The work deals with the issue of recommender systems, namely those that focus on the integration of contextual aspects. Our aim is to improve the recommendation systems by incorporating temporal aspects as one of the important attributes in generating recommendations of Points-of-Interest in domain of location-based social networks.

To prove that, we have proposed methods that addresses long-term trends in popularity of categories of Points-of-Interest and seasonality of categories according to geographical area.

We use temporal modeling enhanced by item features matrix, a temporal post-filter and a temporal pre-filter and a geographical post-filter method based on previous reviewed POIs by a user. The base method is a collaborative-based matrix factorization.

The current results from work with Yelp dataset suggest, that incorporating long-term trends does not seem to be a promising area of research for increasing the accuracy, but the results of local seasonality are significant. They show that implementing of geographical and temporal context as local seasonality improves precision and recall, if weather of the considered areas are significantly influenced by changing seasons. The best results were achieved by contextual modeling.

Obsah

1	Úvod	1
2	Odporúčacie systémy	3
2.1	Typy odporúčacích systémov	4
2.2	Typické problémy odporúčacích systémov	6
2.3	Vyhodnotenie odporúčaní	7
2.4	Diskusia	9
3	Odporúčanie zohľadňujúce kontext	11
3.1	Metódy odporúčania s kontextuálnymi prvkami	12
3.2	Temporálne odporúčanie	16
3.3	Prehľad odporúčacích systémov zohľadňujúcich temporálne aspekty . . .	18
3.4	Diskusia	22
4	Odporúčanie bodov záujmu	23
4.1	Typické vlastnosti odporúčania POI	24
4.2	Aktuálne prístupy kontextuálneho odporúčania v oblasti LBSN	25
4.3	Analýza dostupných datasetov v doméne LBSN	31
4.4	Diskusia	33
5	Návrh metód odporúčania na základe sezónnosti a dlhodobých trendov špecifických pre lokalitu	35
5.1	Konceptuálny návrh metódy	36
5.2	Výber a spracovanie dát	37
5.3	Základné odporúčanie	39
5.4	Temporálne modelovanie	40
5.4.1	Sezónnosť	41
5.4.2	Dlhodobé trendy	42
5.5	Temporálny po-filter	43
5.5.1	Sezónnosť	43
5.5.2	Dlhodobé trendy	43
5.6	Geografický po-filter	43
5.7	Vyhodnotenie	44
5.8	Sumarizácia navrhutej metódy	45
6	Výber a opis dátovej sady pre overenie navrhnutých metód odporúčania	47
6.1	Výber datasetu	47

6.2	Opis datasetu	47
6.2.1	Používatelia	48
6.2.2	Podniky	48
6.2.3	Recenzie	50
6.3	Závislosť dát od časových aspektov	51
6.3.1	Cyklické temporálne vplyvy	51
6.3.2	Necyklické temporálne vplyvy	53
6.4	Diskusia	54
7	Overenie navrhnutých metód odporúčania s temporálnymi aspektami	57
7.1	Realizácia metódy	57
7.1.1	Spracovanie datasetu a výber podmnožiny dát	59
7.2	Metodológia experimentálneho overenia	61
7.2.1	Zapracovanie geografických aspektov	61
7.2.2	Temporálne modelovanie lokálnej sezónnosti s temporálnym pred-filtrom a geografickým po-filtrom	63
7.2.3	Temporálne modelovanie lokálnych trendov s geografickým po-filtrom	64
7.2.4	Temporálny po-filter lokálnej sezónnosti s temporálnym pred-filtrom a geografickým po-filtrom	66
7.2.5	Temporálny po-filter lokálnych trendov s geografickým po-filtrom	67
7.2.6	Finálny výber metódy a jej overenie	67
7.3	Diskusia	72
8	Zhodnotenie	75
	Literatúra	76
A	Plán práce na riešení projektu	A-1
B	Prehľad aktuálnych riešení temporálnych odporúčacích systémov	B-1
C	Ukážka implementácie odporúčania s knižnicou LightFM	C-1
D	Tabuľky výsledkov pre jednotlivé črty modelovania na dátovej sade B	D-1
E	Článok prezentovaný na študentskej vedeckej konferencii IIT.SRC 2019	E-1
F	Technická dokumentácia	F-1
G	Opis digitálnej časti práce	G-1

1 Úvod

Odporúčacie systémy sú momentálne dôležitou súčasťou webových služieb. Či už ide o elektronický obchod alebo prehrávanie videí, práve dobré odporúčania môžu ovplyvniť používateľove správanie a pomôcť udržať si zákazníka. Mnoho súčasných spoločností si to uvedomuje, no stále sa zabúda na to, že aj zohľadňovanie kontextu v odporúčaní môže byť silnou zbraňou a systémy naň častokrát neberú ohľad.

My sme sa rozhodli zamerať na oblasť odporúčania na základe časových aspektov, keďže tie vo veľkej miere môžu ovplyvniť preferencie používateľa. Navrhujeme metódu, ktorá zohľadňuje lokálny temporálny vplyv pri odporúčaní. Metódu overujeme pomocou experimentov nad dátovou sadou zo systému Yelp.

V Kapitole 2 opisujeme základné princípy odporúčania, zatiaľ čo v Kapitole 3 približujeme význam jeho rozšírenia o kontextuálne vplyvy. Problematiku odporúčania na základe času opisujeme v Kapitole 3.2 a Prílohe B a ich praktickú aplikáciu v Kapitole 3.3. Doménu sociálnych sietí založených na lokalite sme bližšie rozobrali v Kapitole 4, rovnako ako aj prehľad aktuálnych prác v tejto oblasti.

Ako nevyriešené problémy v oblasti odporúčania bodov záujmu sme vyhodnotili preskúmanie závislosti sezónnosti od lokality a dlhodobých trendov v popularite kategórií bodov záujmu od lokality. V Kapitole 5 navrhujeme metódu, ktorá vykonáva geografické po-filtrovanie, temporálne po-filtrovanie i modelovanie, v prípade sezónnosti aj pred-filtrovanie. Základom odporúčania je faktorizácia matíc na základe kolaboratívnych princípov.

V Kapitole 6 sme analyzovali dáta z na lokalite založených sociálnych sietí Yelp. Kapitola 7 opisuje konkrétne experimenty a aktuálne výsledky navrhutej metódy na tejto dátovej sade.

Príloha A obsahuje zhodnotenie plánu práce počas semestrov. Príloha B obsahuje prehľad prác zaoberajúcich sa temporálnym kontextom. Ďalšia Príloha 8 obsahuje ukážku implementácie odporúčania pomocou knižnice LightFM a Príloha D zase výsledky jej použitia pri modelovaní temporálnych vplyvov pomocou črt. Časť práce je opísaná aj v článku zo študentskej vedeckej konferencie (Príloha E). Ako posledné dokument obsahuje technickú príručku a samotný opis digitálnej časti práce (Prílohy F a G).

2 Odporúčacie systémy

Softvérové nástroje a techniky, ktoré sa používajú na poskytovanie návrhov položiek (odporúčané objekty), ktoré sú najviac pravdepodobne preferované určitým používateľom, sa nazývajú odporúčacie systémy (ang. *recommendation system*, skrátené RS) [48, 9, 46]. Tento prístup sa dá použiť na početných doménach, či už sa jedná o odporúčenie čo si má používateľ kúpiť, aký film pozrieť či čo si prečítať.

Jeden odporúčací systém sa spravidla zameriava práve na jeden druh položiek - teda odporúča len filmy, len články na prečítanie alebo len predmety pre nákup. Pritom odporúčacia technika je prispôbena pre poskytnutie užitočných a efektívnych návrhov pre práve tento jeden špecifický druh produktov.

To ale neznamená, že položky musia byť odporúčané len ako jednotlivé prvky, je možné odporúčať aj ich súbor, kolekciu [47], ktorej prvky k sebe dobre pasujú. Toto je bežné napríklad v turistickom priemysle, kde v rámci návštevy mesta chce používateľ navštíviť viacero atrakcií.

Výsledok odporúčania [48, 38, 4] si môžeme predstaviť vo forme zoznamu položiek, pričom každej z nich je priradená hodnota na základe predpokladanej užitočnosti, a práve podľa tejto hodnoty sú v zozname zoradené. Hodnotenia položiek vznikajú na základe istých predpokladov, ktoré pramenia zo zozbieraných dát, napríklad o doterajších systémom zaznamenaných činnostiach (transakciách) a vyjadrených preferenciách používateľov (hodnotenia produktov). Informácie môžu ale prameniť aj z vedomostí o konkrétnych položkách alebo ich kontextu.

S rozmachom odporúčacích systémov sme sa mohli stretnúť najmä v posledných rokoch. Základnými dôvodmi podľa autorov publikácie [48] pre to sú:

- zvýšenie predaja,
- predaj väčšej diverzity produktov,
- nárast spokojnosti zákazníkov,
- zabezpečenie vernosti zákazníkov,
- lepšie pochopenie preferencií zákazníkov.

Najčastejšie domény, pre ktoré sa odporúčacie systémy využívajú, vieme rozdeliť do niekoľkých základných kategórií [48]:

- zábava (filmy, hudba atď.),
- obsah (personalizované noviny, emailové filtre atď.),

- e-commerce (odporúčanie produktov na zakúpenie, napr. knihy),
- služby (napr. cestovné služby, odporúčanie prenajatia konkrétnej nehnuteľnosti),
- sociálne médiá (odporúčania ohľadom sociálnych sietí, napr. tweety).

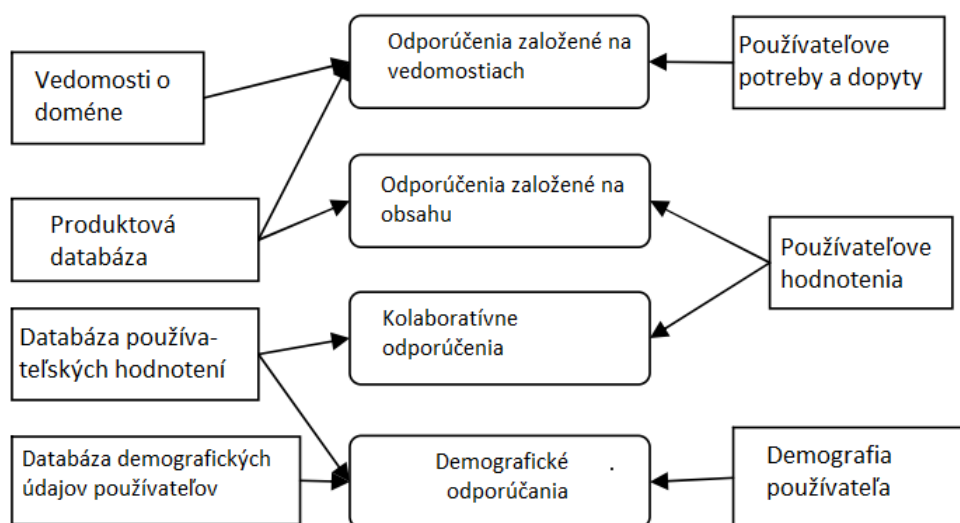
Medzi známe spoločnosti využívajúce odporúčania patrí napríklad Amazon.com (knihy a mnoho iných produktov), Netflix (filmy, seriály, videá), GroupLens (správy), MovieLens (filmy), last.fm (hudba), Google (správy, reklamy), Facebook (spojenia, reklamy), Pandora (hudba), YouTube (videá), IMDb (filmy) či Tripadvisor (cestovateľské produkty).

Pre jednotlivé domény sa využívajú rôzne metódy odporúčania, vybrané na základe množstva faktorov.

2.1 Typy odporúčacích systémov

Spôsoby odporúčania, založené na zdroji vedomostí, vieme zaradiť do štyroch tried [9, 3, 48] (Obrázok 1):

- **kolaboratívne** (ang. *collaborative filtering*, hľadá používateľov s podobnými hodnoteniami a vyberá z nimi pozitívne hodnotených produktov),
 - založené na pamäti (ang. *memory-based*, na susednosti založené kolaboratívne filtrovanie),
 - * založené na používateľoch (ang. *user-based*, hľadá používateľov s podobnými hodnoteniami položiek),
 - * založené na položkách (ang. *item-based*, je založené na podobnosti medzi položkami vypočítanou pomocou hodnotenia používateľov týchto položiek),
 - založené na modelovaní (ang. *model-based*, používa sa strojové učenie a doľovanie dát v kontexte prediktívnych modelov),
- **založené na obsahu** (ang. *content-based filtering*, berie v úvahu atribúty predchádzajúcich hodnotených produktov),
- **demografické** (ang. *demographic*, rozhoduje na základe demografického profilu používateľa),
- **založené na vedomostiach** (ang. *knowledge-based*, berie v úvahu explicitne dané používateľove potreby a preferencie, atribúty položiek a doménovú znalosť),
- **hybridné** (ang. *hybrid*, kombinácia rôznych metód).



Obr. 1: *Techniky odporúčania a ich zdroje vedomostí [9].*

Je niekoľko spôsobov, ktorými rátať blízkosť položiek a používateľov pri **kolaboratívnom prístupe**, medzi najčastejšie používané patria Kosínusová a Pearsonova vzdialenosť. Všeobecne sa považuje metóda založená na položkách za stabilnejšiu, a to najmä z dôvodu, že zvyčajne systém obsahuje viac používateľov ako položiek, preto je vyššia pravdepodobnosť nájsť dve položky s vyšším počtom používateľov, ktorí ich obe ohodnotili, ako nájsť dvoch používateľov, ktorí hodnotili veľké množstvo rovnakých položiek. Preto aj jedno hodnotenie pri metóde založenom na používateľoch môže spraviť oveľa väčší rozdiel, než pri metóde založenej na položkách.

Všeobecnými výhodami metód využívajúcich susednosť sú:

- opodstatnenosť,
- efektívnosť,
- stabilita.

Čo však môže kolaboratívne systémy negatívne ovplyvniť sú časté prvotné fázy testovania a vyvíjania nad offline dátami (ak napríklad testujeme nad offline dátami, nevieme, či používateľ nekúpil položku, lebo o ňu nemal záujem, alebo mu ani nikdy nebola ponúknutá) či riedkosť dát (čím viac dát, tým presnejšie sa dá vyhodnotiť vzdialenosť susedov).

Naopak, **systémy založené na vedomostiach** sú užitočné najmä pri nákupoch, ktoré sú veľmi občasné (napríklad nákup nehnuteľnosti, auta, luxusného tovaru).

Existujú však aj systémy využívajúce viacero riešení súčasne, tzv. **hybridné systémy** [8]. Používajú rôzne metódy, akými kombinujú jednotlivé odporúčacie prístupy:

- váhovanie (ang. *weighted*, skóre viacerých odporúčacích komponentov kombinované numericky),
- prepínanie (ang. *switching*, v rôznych situáciách sa použije rôzna metóda odporúčania),
- zmiešané (ang. *mixed*, odporúčania viacerých metód sú prezentované spoločne),
- kombinácia vlastností (ang. *feature combination*, vlastnosti z rôznych znalostných zdrojov sú spolu skombinované a dané jednej odporúčacej metóde),
- obohatenie vlastností (ang. *feature augmentation*, odporúčacia technika použije vlastnosti, ktoré vyprodukovala iná metóda),
- kaskáda (ang. *cascade*, systémy sú zostavené hierarchicky, menej dôležité rozhodujú remízy v odporúčaní vyššie postavených),
- meta-úrovňové (ang. *meta-level*, technika odporúčania využije čiastočný model, ktorý vyprodukovala iná metóda odporúčania).

Ďalším prístupom, ktorý využívajú najmä kolaboratívnej metódy, je **faktorizácia matic** [3, 48] (skrátene MF z ang. *matrix factorization*). Tá slúži na rozloženie matíc (používateľa x položky, kde ich prienik je napríklad hodnotenie položky používateľom alebo inak vyjadrené preferencie), konkrétne z jednej matice spraví dve. Tieto dva komponenty vzájomným vynásobením následne vytvoria opäť pôvodnú maticu. Tento postup je nápomocný pre zmenšenie príznakového priestoru kombináciou pôvodných atribútov. Takto vzniká nová sada vlastností, ktoré popisujú dané objekty, a v konečnom dôsledku pomáha dosiahnuť lepšie výsledky. MF sa momentálne považuje za *state-of-the-art* metódu, teda modernú špičkovú metódu v oblasti odporúčacích systémoch.

2.2 Typické problémy odporúčacích systémov

Pre odporúčacie systémy sú charakteristické viaceré problémy. Jedným z nich je studený štart (ang. *cold-start*) [9]. Podstata problému [3] spočíva v nedostatku informácií o novom používateľovi alebo položke, ktoré sú pre techniky, založené na učení, nevyhnutné.

Medzi ďalšie problémy patria tzv. čierne a sivé ovce (ang. *black sheep* a *grey sheep*) [65]. Prvú skupinu tvoria používatelia, ktorí majú málo alebo žiadnych korelujúcich používateľov. Ako sivé ovce sú označovaní používatelia, ktorí majú rozdielne názory alebo neobvyklý vkus v porovnaní s väčšinou, a teda ich korelácia s inými používateľmi je veľmi nízka.

Okrem toho je pri kolaboratívnom odporúčaní často problematická aj distribúcia hodnotení medzi položkami, ktorá vytvára tzv. dlhý chvost, ang. *long tail* [3]. Znamená to,

že len malá časť položiek je hodnotená pravidelne, s vysokým počtom hodnotení (tzv. populárne položky), zatiaľ čo väčšia časť položiek je hodnotená len zriedka.

Netreba zabúdať ani na škálovateľnosť odporúčacích systémov či potrebu rozhodnúť, ako často odporúčací model pretrénovať [38].

Z dôvodu časového vývoja nastáva aj ďalšia komplikácia, a to zmena konceptu v čase (ang. *concept drift*) [50].

2.3 Vyhodnotenie odporúčaní

Aby odporúčacie systémy vedeli generovať personalizované odporúčanie pre používateľa, potrebujú mať o ňom isté vstupné informácie, napr. demografické údaje alebo záznamy jeho predchádzajúcich transakcií. Záznamy transakcií slúžia ako spätná väzba, napr. pre kolaboratívne filtrovanie. Príklad: Používateľ si nedávno pozrel niekoľko akčných filmov z posledného roku a tak na základe podobnosti položiek, teda filmov, vieme odporučiť ďalší akčný film. Zber spätnej väzby je preto pre tento typ odporúčania kľúčový, pričom môže prebiehať:

- explicitne (výslovne vyžiadaná spätná väzba, napr. dotazník),
- implicitne (skryto získaná spätná väzba, napr. opakované zakúpenie produktu).

Po vygenerovaní odporúčaní, ich vyhodnotenie môže prebiehať na dátach [10, 3]:

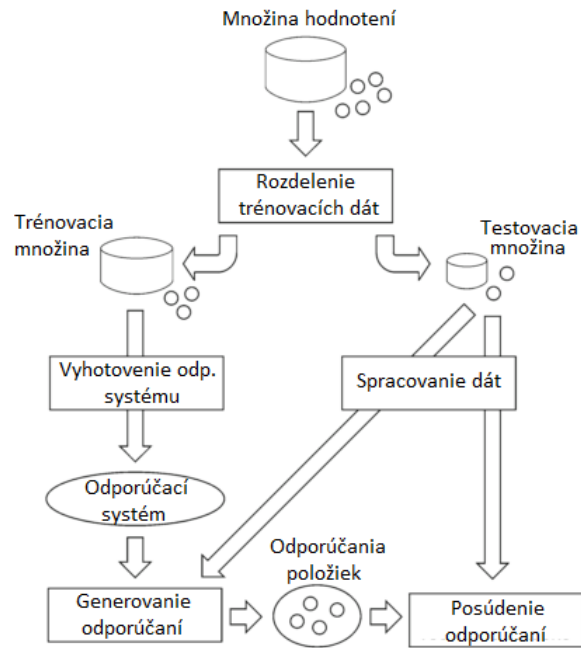
- online,
- offline.

Online znamená vyhodnocovanie priamo v prevádzke, s reálnymi používateľmi, a teda zrkadlí reakcie používateľa s ohľadom na predložené odporúčania, zatiaľ čo offline (Obrazok 2) na množine dát, transakcií, ktoré rozdelíme na množinu pre vývoj odporúčacieho systému a na množinu pre jeho overenie. Množina na overenie, ktorá je spravidla menšia, simuluje správanie používateľov v novom systéme (ang. *ground truth*) [10].

Častokrát je online testovanie sprevádzané A/B testami, teda nový odporúčač sa testuje len na časti používateľov, napríklad polovici. Následne sa porovnávajú výsledky týchto dvoch skupín, aby sa dalo objektívnejšie pozorovať, či prípadný skutočný úspech alebo neúspech je naozaj spôsobený novým odporúčaním.

Pritom sa pri posudzovaní úspešnosti odporúčaní dbá najmä na [17]:

- relevantnosť,
- novosť (odporúča sa čo používateľ ešte nevidel),



Obr. 2: Priebeh odporúčania nad offline dátami [10].

- prekvapivosť (cena sa najmä na prvý pohľad nezrejme riešenia),
- zvýšenie rôznorodosti odporúčania (odporúčené položky by nemali byť príliš podobné).

Keď sa neberie ohľad na posledný bod, rastie pravdepodobnosť, že sa používateľovi nebude páčiť ani jedna z možností. Ak sa dbá na diverzitu, je vysoká šanca, že používateľovi sa bude pozdávať aspoň jedna položka.

Dôležité sú však aj ostatné aspekty. Niektoré práce [3] udávajú medzi podstatnými faktormi aj pokrytie (ang. *coverage*), robustnosť, stabilitu, dôveryhodnosť, škálovateľnosť a iné.

Na meranie týchto žiadaných kvalít systému sa používa niekoľko metrík [3]:

- správnosť (ang. *accuracy*, podiel správnych predikcií),
- presnosť (ang. *precision*, pozitívna prediktívna hodnota),
- úplnosť (ang. *recall*, pravdepodobnosť detekcie),
- skóre F1 (správnosť testu),
- stredná kvadratická chyba (ang. *mean squared error*),
- stredná absolútna chyba (ang. *mean-absolute-error*),

- diskontovaný kumulatívny zisk (ang. *discounted cumulative gain*) alebo jeho normalizovaná hodnota - skrátené NDCG,
- koreňová stredná kvadratická odchýlka (ang. *root-mean-square deviation*, inak nazývaná aj odmocnená stredná kvadratická chyba - ang. *root mean squared error*),
- stredné recipročné poradie (ang. *mean reciprocal rank*, skrátené MRR),
- a iné.

2.4 Diskusia

V tejto kapitole sme poukázali na zmysel odporúčacích systémov ako prostriedok pre zvýšenie spokojnosti používateľov i druhej strany. Tiež sme vysvetlili ich základné princípy, delenia, potencionálne úskalia i spôsoby merania ich úspešnosti. Najviac do hĺbky sme sa venovali kolaboratívne prístupu, najmä v kombinácii s faktorizáciou matíc.

Úspešnosť odporúčacích systémov je vždy možné zvyšovať a jedným z prístupov, ktorým sa dá odporúčací systém posunúť ďalej, je zakomponovanie kontextu medzi posudzované atribúty. Tejto oblasti sa venujeme v Kapitole 3.

3 Odporúčanie zohľadňujúce kontext

Mnoho odporúčacích systémov berie vo svojej metóde najväčší dôraz na dve entity, používateľov a položky, už menší na iné vplyvy [2]. Vyhodnotenie odporúčaní [3, 1, 10] je teda funkciou dvoch premenných (Rovnica č. 1, kde U predstavuje používateľa a I položky, spolu vytvárajú dvoj-dimenziálny priestor):

$$h_R : U \times I \rightarrow \text{hodnotenie} \quad (1)$$

Pritom informácie o kontexte transakcií, ako napríklad s kým si ide používateľ vybrať film pozrieť, môže konečné rozhodnutie značne ovplyvniť. Niektoré spoločnosti tieto spojenia už pochopili a berú do úvahy napríklad náladu pre hudobné odporúčania. Uvedomili si, že je to príležitosť obohatiť interakciu medzi svojimi používateľmi a systémom, spraviť ju viac rozmanitou [1].

Kontext je pritom viacúčelový koncept, ktorý používajú rôzne domény [2, 48] (vyhľadávanie informácií, dolovanie dát, marketing, e-commerce personalizácia, databázy atď.). Jazykovedná definícia označuje kontext ako

okolnosť, že niečo s niečím súvisí a navzájom sa podmieňuje a ovplyvňuje, vzájomný vzťah a vnútorná spätosť medzi určitými javmi, situáciami a pod., súvis, súvislosť¹

pričom význam v iných doménach sa môže zúžiť, resp. rozšíriť. V oblasti odporúčacích systémov sa rozlišuje niekoľko dimenzií kontextu, napríklad:

- čas,
- lokalita,
- spoločnosť ďalších ľudí,
- účel nákupu,
- a iné.

Tieto dimenzie sa následne môžu ešte hierarchicky deliť [3], napríklad čas na hodiny, dni, týždne a tak ďalej.

Samozrejme, nie v každom odporúčaní musia byť obsiahnuté kontextuálne informácie, alebo niektoré ich dimenzie, užitočné [2]. Taktiež, niekedy je lepšie sústrediť sa na viacero dimenzií súčasne (napríklad na čas a lokalitu) [3].

¹<http://slovníky.juls.savba.sk/?w=kontext>

3.1 Metódy odporúčania s kontextuálnymi prvkami

Systémy, ktoré zakomponovali kontext do svojho rozhodovania, pozmenili teda Rovnicu č. 1 na Rovnicu č. 2 pridaním tejto novej zložky [1, 10]:

$$h_R : U \times I \times C \rightarrow \text{hodnotenie} \quad (2)$$

kde okrem U (používatelia) a I (položky) reprezentuje C kontext.

Hoci veľa odporúčacích systémov kontext vôbec neberie do úvahy, aj keď by to pre ne mohlo byť v istých podmienkach prospešné, existuje aj opačný extrém, relatívne nové hnutie, ktoré sa venuje odporúčaniam nie len s ohľadom na kontext, ale berie ho ako svoj základ [40, 2]. Toto hnutie sa nazýva kontextuálny zvrät (ang. *contextual turn*) a čelí niekoľkým výzvam:

- spojitý kontext (opak kategorického kontextu, faktorizačné metódy ho nie sú schopné spracovať),
- chýbajúce hodnoty,
- škálovateľnosť,
- vyhodnotenie.

Čo sa týka zberu počiatočných informácií pre odporúčania založené na kontexte, môže prebiehať [2, 48]:

- explicitne - priamou spätnou väzbou, napr. pomocou dotazníkov,
- implicitne - zo získaných dát, napr. logov konkrétného používateľa,
- používaním štatistických metód alebo skúmaním dát - získanie potrebných informácií pomocou predikcie a klasifikácie.

Okrem toho je možné, aby kontextuálne informácie boli v dátach v latentnej podobe, teda použiteľné na odhad neznámych hodnotení bez konkrétnej informácie o kontexte.

Pritom odporúčací systém môže mať rôzne vedomosti o kontextuálnych faktoroch [48, 21], ktoré sa delia na:

- plne pozorovateľné,
- čiastočne pozorovateľné,
- nepozorovateľné (implicitné latentné vedomosti).

Všetky môžu pritom byť statické (stabilné v čase) alebo dynamické (meniace sa).

Existuje niekoľko spôsobov, ako následne modelovať kontextuálne informácie pri odporúčaní (Obrázok 3):

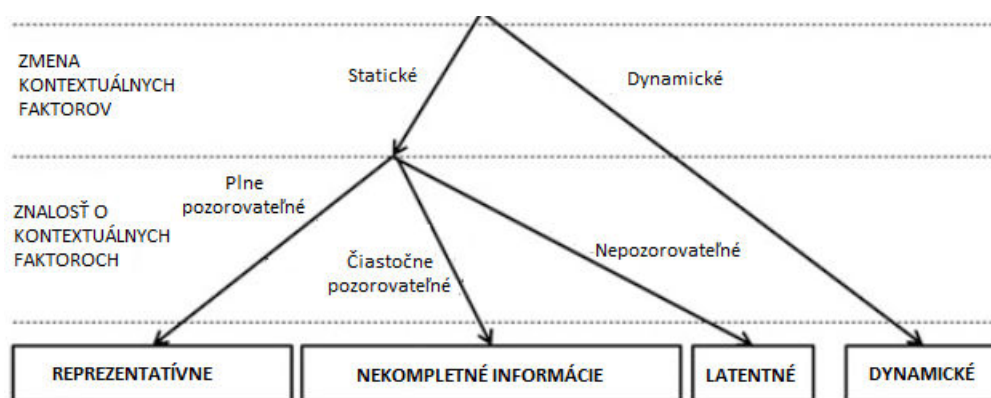
- reprezentatívny,
- nekompletné informácie,
- latentný a
- dynamický.

Reprezentatívny spôsob predpokladá, že všetky kontextové informácie v danej aplikácii môžu byť modelované s vopred definovanou konečnou množinou pozorovateľných atribútov, kde každý atribút má dobre definovanú štruktúru a štruktúra sa časom nemení.

Prístup nekompletných informácií explicitne predpokladá, že sú o faktoroch známe len čiastočné informácie a tie sa považujú za statické.

Latentný prístup považuje faktory za statické, ale priamo nepozorovateľné. Napríklad takéto kontextuálne faktory môžu zahŕňať zámer nákupu (pre seba, pre prácu, darček atď.). Z tohto dôvodu môže byť ťažké diferencovať kontextové latentné modelovanie z všeobecne používaných latentných modelovacích prístupov v odporúčaných systémoch (napr. prístupy faktorizácie matíc, ktoré sú populárne). Rozlíšiteľné je to na základe možného prepojenia latentných premenných s položkou, užívateľom a/alebo kontextovými charakteristikami.

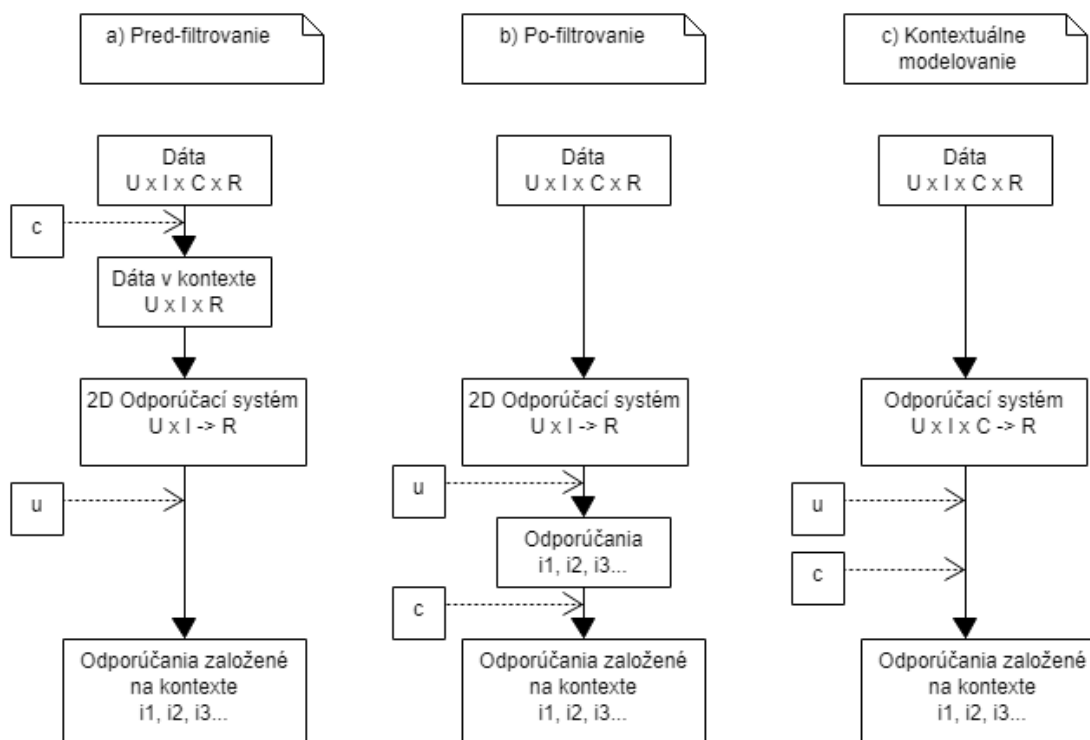
Posledný prístup, dynamický, zastrešuje prípady, kde sa kontextuálne informácie menia naprieč časom.



Obr. 3: Najčastejšie prístupy v modelovaní kontextuálnych informácií [48].

Tieto informácie sú následne využívané metódami na zapracovanie kontextu, ako sú [2, 3] (Obrázok 4):

- Pred-filtrovanie - segment hodnotení je filtrovaný vopred podľa relevantnosti kontextu, až potom je táto vybraná časť využitá na vytvorenie cieľového odporúčania.
- Po-filtrovanie - hodnotenia sú najprv vygenerované v rámci všeobecnej metódy odporúčania, až následne sú výsledky preusporiadané aj na základe kontextu.
- Modelovanie - informácie o kontexte sú využívané počas celého procesu odporúčania, nevyužíva teda dvojdimenzionálny odporúčací systém, je obzvlášť výhodné pre veľký objem dát, na výpočet je ale najnáročnejšie.



Obr. 4: Metódy odporúčania aj na základe kontextu [2], kde U/u sú používatelia, I položky (jednotlivá položka i), C/c kontext a R odporúčania.

Výhodou pred-filtrovaní [2], ktoré sa niekedy nazýva aj redukcia, je možnosť výberu z rôznych rozvinutých metód odporúčania, ktoré sa používajú aj pre bezkontextové systémy a môžu sa použiť aj v tomto prípade, pričom kontext môže byť použitý pre nájdenie podstatných dát, z ktorých sa následne dá ťažiť (napr. ak niekto ide do reštaurácie cez víkend, systém sa pozrie len na víkendové hodnotenia reštaurácií). Má to však aj svoje úskalí, a to najmä ako správne rozdeliť dáta do skupín kontextu - to, že niekto ide do reštaurácie v sobotu, nemusí znamenať, že v piatok večer by sa zachoval inak. Nie všetky aspekty špecifického kontextu budú významné. Problémom môže byť tiež nedostatok dát,

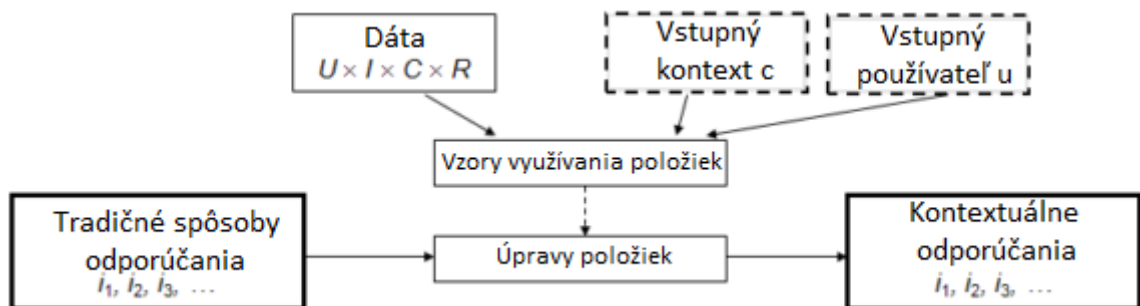
pokiaľ sa ich rozhodneme takto rozdeliť do skupín. S týmito problémami by ale mohol pomôcť expertný prístup.

Pri po-filtrovaní sa tiež pomocou niektorého z bežných odporúčacích algoritmov vyberú najvhodnejšie položky, ale nepredchádza tomu filtrovanie dát na základe kontextu. Zmeny spraví až na tomto zozname, buď vynechaním položiek, ktoré sú v danom kontexte nepodstatné, alebo zmenou ich poradia, tiež podľa informácií z relevantného kontextu. Priebeh je možné vidieť na Obrázku č. 5. Tento prístup môže využiť dve techniky, buď zvoliť:

- heuristiku (nájdenie spoločných atribútov v danom kontexte) alebo
- modelovanie (budovanie prediktívnych modelov).

V prípade heuristiky sa následne atribúty používajú na:

- vyradenie položiek, ktoré nemajú dostatočný počet z týchto atribútov,
- zoradiť položky podľa počtu týchto hľadaných atribútov.



Obr. 5: Priebeh po-filtrovaní [2].

Podobne pristupuje aj modelovanie pri po-filtrovaní, ktoré:

- vyradí položky, ktorých pravdepodobnosť výberu je menšia ako zadaná hranica,
- zoradiť položky podľa predpokladaného hodnotenia.

Posledným prístupom je kontextuálne modelovanie [2, 48], ktoré umožňuje multi-dimenzionálne odporúčanie v podobe prediktívnych modelov alebo heuristické výpočty. Heuristický prístup môže mať podobu napríklad klasického modelu založeného na vzdialenostiach susedov, ktorý sa zmení na multidimenzionálny pridaním kontextuálnych informácií. Podobnosť sa ale v tomto prípade vníma ako blízkosť metrík v multidimenzionálnom priestore.

Nedá sa jednoznačne povedať, že jeden prístup je lepší, ako iný [48]. Modelovanie a po-filtrovanie dosahujú dobrú správnosť, ale pred-filtrovanie zase dosahuje lepšiu rôznorodosť. Avšak, modelovanie vykazuje celkom dobrú rovnováhu medzi správnosťou a diverzitou, ale vo vysoko-dimenzionálnom prostredí môže byť výpočtovo priveľmi náročné. Samozrejme, vždy je možné využiť aj kombináciu viacerých prístupov. Dá sa spojiť aj niekoľko prístupov s rôznymi aspektami kontextu do jedného konečného odporúčania.

3.2 Temporálne odporúčanie

Jednou z často využívaných kontextuálnych dimenzií je čas [3]. Ten dokáže v niektorých doménach hrať rozhodujúcu úlohu, napríklad ako vek položky - iný je prístup k filmu tesne po premiére a iný niekoľko rokov po jeho vydaní. To isté platí vo viacerých oblastiach, keďže mnohé inštancie (filmy, hudba, hry atď) majú tendencie podliehať časovo obmedzeným trendom. Tiež dobrý príklad sú domény, v ktorých sa objavujú silné sezónne vzory [40], napríklad pravidelné pozeranie nejakého TV kanálu alebo činnosti v oblasti turizmu. V tom sa dokonca ukazujú v dvoch rôznych podaniach, a to absolútnom kontexte (je práve hlavná sezóna?) a relatívnom kontexte (koľko dní vopred si používateľ rezervuje ubytovanie?). Dôležité sa dokonca ukazuje nie len čo odporúčať na základe času, ale aj kedy [53].

Branie v úvahu časových aspektov je veľmi náročné z dôvodu potreby vysokého objemu dát pri vytváraní systému, avšak je dokázané [2], že venovať sa tejto oblasti odporúčaní má zmysel a prístup s ohľadom na temporálne aspekty, v kombinácii s tradičnými (používatelia, položky) dokáže systém zlepšiť.

Prejavy časových aspektov pritom môžeme rozdeliť do dvoch kategórií [3, 6]:

- modely založené na aktuálnosti - pretože hodnotenia produktu, preferencie komunity či záujmy používateľa sa menia v čase, alebo
- periodické modely - hodnotenia môžu byť závislé na špecifickom období počas dňa/týždňa/mesiaca/ročného obdobia/podobne, ktoré sa zvyknú opakovať.

Ako objekt, vzhľadom na ktorý sa môžu sledovať časové aspekty, slúži:

- používateľ alebo
- položka

Ak používateľ počas dospievania preferoval filmy-komédie, neznamená, že ich preferuje aj o niekoľko rokov neskôr, rovnako nákup lyžiarskych športových potrieb v decembri neznamená, že o ne bude mať záujem aj počas leta.

Čo sa týka spôsobov, akými môžeme reprezentovať čas, jedná sa o nasledovné [10]:

- reprezentovať čas ako spojitú kontextuálnu premennú,
- ako kategorickú kontextuálnu premennú,
- pomocou časovo adaptívneho prístupu (využíva časové informácie bez diferenciácie predikcie hodnotenia podľa cieľového času, ale skôr dynamické prispôsobenie niektorých parametrov alebo údajov).

Pritom časové aspekty sa dajú získať dvomi rôznymi spôsobmi [3]:

- explicitné hodnotenie (dátumy spojené s explicitnými hodnoteniami),
- implicitná spätná väzba (zodpovedajú činnostiam zákazníka, ako nákup položiek/kliknutie na webové stránky/atď.).

Čas môžeme použiť pre zistenie novosti (novšie hodnotenia sa pokladajú za dôležitejšie), predikciu alebo kontextuálnu perspektívu. Temporálne aspekty sa teda môžu chápať v troch rôznych rovinách, ako periodické, s pohľadom na novosť alebo ako modelovanie.

Jednou možnosťou je teda spraviť model založený na aktuálnosti. Pod takýmto modelom si môžeme predstaviť napríklad doménu novinových článkov, pri ktorých je nesmierne dôležité uprednostňovať najnovšie správy, takže novším hodnoteniam sa dáva väčší dôraz ako starším. Tento prístup sa ďalej delí na:

- modely založené na rozklade (hodnoteniam sa postupne znižuje dôležitosť podľa funkcie úpadku),
- modely založené na aktuálnosti (špeciálny prípad rozkladu, hodnotenia staršie ako daný čas sa vôbec neberú v úvahu).

Okrem modelov založených na aktuálnosti sú tu aj modely založené na periodicite, ktoré sú založené na špecifickej vlastnosti periódy (napr. špecifický deň, týždeň, ročné obdobie). Príkladom môžu byť filmy, ktorých odporúčenie by sa malo líšiť na základe toho, či sú Vianoce alebo obdobie okolo udeľovania Oscarov - v takomto prípade sa stane čas kontextuálnou premennou.

Poslednou možnosťou je, aby čas bol priamo premennou, ak sú hodnotenia funkciou času. Táto funkcia môže vzniknúť minimalizáciou kvadratickej chyby predikovaných hodnotení k pozorovaným hodnoteniam pri dátovo riadenom prístupe. Príkladom môže byť časová singulárna dekompozícia (ang. *time singular-value decomposition*, skrátene *time-SVD++*), čo je v lineárnej algebre faktorizácia reálnej alebo komplexnej matice. Vyjadruje hodnotenia ako funkciou temporálnych parametrických odchýlok a faktorových matíc. Následne dokáže odhaliť budúci vývoj, pričom využíva vylepšené trendy špecifické pre

používateľa a špecifické pre jednotlivé položky na zvládnutie menších časových zmien a môže tiež zohľadňovať prerušovaný časový šum v hodnoteniach.

Zatiaľ čo metóda SVD++ vie zachytiť širšiu škálu charakteristík, zvyšné dva modely je oveľa jednoduchšie naimplementovať.

Typickým problémom odporúčaní berúcich ohľad na temporálne aspekty je v Kapitole 2 spomínaný *concept drift* [7]. Je to stav, pri ktorom sa používateľove záujmy významne zmenia v priebehu času. Sú pritom rozlišované jeho tri druhy a riešenia [52]:

- výberu inštancie (podstatné pre súčasné časové okno),
- vážení inštancie (na základe odhadovanej podstatnosti) a
- učenia súboru (zaoberá sa prediktormi, váženie na základe súčasnej podstatnosti).

Existujú aj prípady, kde sa viacero charakteristík mení súčasne [31], pričom sa ovplyvňujú a často je to spojené len s malým množstvom dát, čo je rozdielne od predchádzajúceho problému. Koren et al. [31] navrhli riešenie v podobe modelu, ktorý sleduje zmeny správania za celú dĺžku životného cyklu dát. To v ich prípade umožnilo odhaliť relevantnosť jednotlivých komponentov inštancií. Použili pritom kolaboratívne filtrovanie a dáta z Netflixu².

3.3 Prehľad odporúčacích systémov zohľadňujúcich temporálne aspekty

V tejto podkapitole analyzujeme aktuálne prístupy v oblasti odporúčacích systémov, ktoré berú do úvahy temporálne aspekty. Ich prehľad je v Prílohe B.

Niekoľko prác sa posledné roky zaoberalo aplikáciou časových vplyvov na odporúčanie. Napríklad v [4] analyzovali výkon dvoch metód, ktoré taktiež brali v úvahu dočasné aspekty používateľského profilu:

- prvá bola zameraná na funkciu časového rozpadu za účelom zdôraznenia dôležitosti najnovších položiek v profile používateľa,
- druhá vykonávala hodnotenia založené na identifikácii a analýze časových relácií.

Podarilo sa im dosiahnuť zlepšenie z hľadiska rovnováhy medzi správnosťou a rozmanitosťou v oboch prípadoch, ale mierne lepšie výsledky dosiahli s prístupom zameraným na temporálnu dynamiku.

²<https://www.netflix.com>

Novinové články

V práci [38] sa tiež museli zamerať na novosť položiek, keďže sa zaoberali doménou novinových článkov, ktorá je typická krátkym životným cyklom položiek a neustálym problémom studeného štartu. Použili niekoľko metód, na konci ich porovnávali na základe súčtu získaných kliknutí používateľov na nimi odporúčané články.

Začali analýzou nad dostupnými dátami, kde skúmali, ako sa správanie používateľov mení s ohľadom na čas, napríklad v priebehu dňa alebo týždňa. Potvrdilo sa, že to podlieha skutočne istým vzorom správania - napríklad dopoludnia boli populárnejšie články o politike, kým večer tie o životnom štýle. Aj to, či väčšina používateľov využila mobilný telefón/tablet/počítač sa v priebehu dňa menilo. Dôležitým prvkom bola však aj odchýlka popularity jednotlivých článkov, ktorá sa menila.

Vyskúšali niekoľko prístupov v akých intervaloch pretrénovať model:

- periodicky - keď sa nazbiera určitý počet článkov,
- nepretržite - vždy keď príde nová správa, najstaršia sa odstráni,
- podľa dňa v týždni/hodiny dňa - v určitých intervaloch,
- krátkodobo na základe extrapolácie - analyzuje poslednú hodinu prúdu dát a identifikuje krátkodobé trendy, používa sa na zachytenie neočakávaných udalostí.

Výsledky porovnávali s výsledkami pri stratégii odporúčaní najpopulárnejších položiek. Každý z využívaných spôsobov má svoje výhody aj nevýhody a rozhodnutie, ktorú metódu využiť, by malo vychádzať z vedomostí ohľadom riešeného scenára, pričom je možné metódy aj kombinovať.

Výhodou periodického prístupu bolo jednoduché vysporiadanie sa s prúdom dát, nevýhodou nebranie v úvahu životného cyklu položky. Pre analýzu vytvorili štyri odporúčacie systémy, ktoré zoraďovali položky podľa rôznych kritérií (náhodne, dátum vydania, počet prečítaní, počet otvorení po článku, ktorý používateľ aktuálne číta). Vždy po vytvorení nového modelu (každých 24 hodín) boli výsledky výrazne lepšie, avšak z dlhodobého hľadiska sa prístup neukázal ako efektívny, keďže v priebehu dňa strácal efektívnosť.

Následne teda prebehol pokus obnovovať modely priebežne, vždy po prijatí novej položky, pričom najstaršia položka sa vylúči. Tento model, založený na posuvnom okne sa ukázal viac úspešný, avšak experiment bol vykonaný nad pomerne malým množstvom dát (300 položiek) a hlavne v praxi nie všetky modely podporujú priebežné aktualizácie.

Ďalší model bol založený na špecifickom správaní používateľov počas rôznych dní týždňa a rôznych hodín počas dňa, teda vytvárania špecifického modelu pre každú hodinu počas dňa. Zakladá si pritom na klasifikácii typov správ. Bolo vybraných niekoľko

odporúčacích algoritmov, každý berúc v úvahu iné atribúty položky, a z nich je počas dňa vyberané na základe predchádzajúcich zistení. Výhodou zakladania si na dlhodobých vzoroch správania používateľov je možnosť trénovania na veľkom objeme dát, ak je k dispozícii. Dajú sa minimalizovať rušivé elementy a spoľahlivo vybrať správny odporúčací systém pre danú hodinu. Problémy ale môžu vzniknúť pri nepravidelných udalostiach (napr. pohyblivé sviatky počas roka), krátkodobých zmenách (napr. mimoriadne správy) či veľmi dlhodobé preferencie (napr. niekoľko rokov).

Posledným skúmaným modelom bolo odporúčanie založené na predpokladanej popularite položky - teda iteratívny model, ktorý by mal zvládnuť aj nepredvídateľné udalosti. Sledujú nové informácie počas posledných pár hodín a identifikujú krátkodobé trendy. Predikujú teda preferencie používateľov, ale aj relevantnosť konkrétnej položky, na základe jej životného cyklu. Počet prečítaní v nasledujúcom časovom okne sa pritom rátať vzťahom

$$impr_{t+1} = impr + 0,8 \times (impr - impr_{t-1}) \quad (3)$$

kde $impr$ znamená počet kliknutí na článkov za posledných 15 minút. Problémom je, že položka má častokrát len malý počet prečítaní na relevantné porovnanie.

S rovnakým datasetom a niekoľkými ďalším pracoval aj iný tím [11]. Snažili sa pritom o zlepšenie rovnováhy medzi aktuálnosťou a relevantnosťou odporúčaných článkov.

Tejto doméne sa venuje viacero prác, napríklad aj demo v [51], ktoré pracuje nad dátach z Outbrain (webová služba, ktorá odporúča veľké množstvo príbehov denne). Kombinujú kontextové stopy s personalizáciou. Aspekty personalizácie sú kombinované s odporúčaním založenom na obsahu a kolaboratívnym odporúčaním.

Črty, ktoré berie v úvahu:

- kontextuálne (aj keď nie temporálne),
- zohľadňujúce popularitu a výkon,
- osobnostné.

Iným datasetom v tejto doméne je Reuters Corpus Volume 1, nad ktorým sa taktiež ukázalo, že novosť položiek je dôležitá [16].

Hudba

Ďalšou doménou, v ktorej sa uplatnilo zohľadniť temporálne aspekty, bola hudba [41]. Záznamy z dvoch rokov zo služby Last.fm sa rozdelili po rokoch na tréningovú a testovaciu množinu a výsledky ukázali zlepšenie odporúčaní o 5 %. Okrem časových vplyvov zohľadňovali aj vplyv používateľov medzi sebou. Pritom za udalosť v systéme

považovali, ak dvaja priatelia počúvali toho istého interpreta krátko po sebe. Ako metódu použili kombinácie prístupov faktorizácie matíc (tá modelovala vzťah používateľ-priateľ_ktorý_ho_ovplyvnil-interpret), pomocou jednotlivých dekompozícií faktorov aj hmotností, naučených variantami stochastického gradientového zostupu.

Odporúčaniu nad hudobnou doménou Last.fm sa venovali aj autori práce [49] vo vytvorení tutoriálu, ako odporúčať hudbu. Navrhli pritom tri scenáre:

- automatické generovanie zoznamu piesní na prehranie,
- odporúčanie v kreatívnom procese tvorby hudby a
- kontextové hudobné odporúčanie.

Ako kontextové vplyvy na odporúčanie hudby udávajú kontext používateľa (napríklad emócie, aktivita, temporálny kontext) a kontext hudby (napríklad video klip, ilustračný obrázok, tagy). Za kľúčové metriky považujú pritom správnosť, rovnováhu novosť-popularita/známosť, tiež diverzita-podobnosť. Inými dôležitými faktormi sú aj transparentnosť a zdôrazňujú význam zohľadnenia kontextu.

Práve dynamike používateľských preferencií sa venovali na rovnakom datasete [44]. Ich úlohou bolo skúmať najmä opakované využívanie hudby.

Nad podobným datasetom a dvomi ďalšími odporúčali hudbu aj [12]. Zamerali sa pritom na heterogénne dopyty (používateľ môže vyhľadávať na základe interpreta, piesne, albumu), ktoré následne porovnávali s výsledkami z kolaboratívneho filtrovania s faktorizáciou matíc (použili knižnicu My-MediaLite³) či výsledkami založenými na odporúčaní populárnych položiek.

Autori [26] tiež, okrem štyroch ďalších, využili aj doménu hudby a dataset z Last.fm. Zamerali sa pritom na modelovanie spojitých kontextuálnych dimenzií, konkrétne pomocou “fuzzy” kontextuálneho modelovania, ktoré umožňuje prekrývanie stavov. Ako príklad kontextu sa používa sezónnosť. Tieto koncepty boli zahrnuté do algoritmu iTALS (faktorizácia s ohľadom na kontext).

Filmy

I iné médiá sú používané ako domény, napríklad filmy [30], taktiež využili faktorizáciu matíc. Taktiež [59] analyzuje používateľské správanie v sociálnych médiách vo vzťahu k filmom. Ako kontext vnímajú najmä vnútorné záujmy používateľov a témy súvisiace s časovým kontextom. [31] modeluje nad dátami z Netflixu aj s faktorizáciou, aj v modeli susednosti položiek, pričom poukazujú na vplyv položiek medzi sebou, resp. ako sa jeho význam postupne časom od hodnotenia znižuje. V niektorých prípadoch sa podobné domény spájajú aj so skupinovým odporúčaním [43].

³<http://www.mymedialite.net/>

Sociálne siete

V neposlednom rade sa temporálne aspekty zohľadňujú v sociálnych sieťach, napríklad Twitter [33]. Niekedy sú kombinované so skupinovým odporúčaním [25].

Medzi zaujímavé skúmané sociálne siete patria siete zamerané na cestovanie [14] alebo inak zohľadňujúce geografické vplyvy.

Všeobecne

V mnohých prácach, ktoré zatiaľ zahŕňajú kontextuálny vplyv, ale bez časových aspektov, sú práve temporálne prvky smerom, ktorým sa chcú autori ďalej vydať [20]. Dokonca aj domény ako integrované vývojové prostredie [19] začínajú zahŕňať kontext v podobe periodických temporálnych faktorov.

Zdá sa, že často používanou metódou je faktorizácia matíc, ktorá je využívaná v kontextuálnom odporúčaní aj mimo temporálnych aspektov [42, 24].

3.4 Diskusia

V Kapitole 3 sme rozšírili princípy odporúčacích systémov oproti vedomostiam z Kapitoly 2 o tie, ktoré berú ohľad na aspekty kontextu. Skúmali sme, aké rôzne dimenzie kontextu sa môžu pri odporúčaní vyskytnúť a ako ho spracovať do odporúčania. Pritom sme sa bližšie zamerali na temporálne aspekty, ktoré predstavujú jeho zložku času.

Preskúmali sme rôzne dimenzie kontextu pri odporúčaní a jeho možnosti spracovania, s dôrazom na konkrétne prípady metód, ktoré zahŕňajú temporálne aspekty. Pri ich analýze sme objavili niekoľko dostupných datasetov, vhodných na vyhodnotenie metód a prístupov v tejto oblasti.

Z možností sme sa rozhodli ďalej sa venovať kontextovému odporúčaniam sociálnych sietí berúcich ohľad na geografické aspekty, ktoré sa objavovali zriedkavejšie ako napr. odporúčanie novinových článkov či filmov, ale pritom ponúkajú unikátne prepojenie temporálneho aspektu s geografickým a mnoho ďalších výziev.

4 Odporúčanie bodov záujmu

Posledné desaťročie sa do popredia dostávajú viac a viac internetové sociálne siete. Jedným z jej druhov sú aj tzv. *Location-Based Social Networks*⁴ (sociálne siete založené na lokalite, skrátene LBSN), kde sa používatelia môžu prihlásiť na konkrétnych reálnych miestach a zdieľať aktivity či skúsenosti vo fyzickom svete prostredníctvom mobilných zariadení [5]. Príkladmi takýchto sietí sú Foursquare⁵, YELP⁶ či už zaniknutá Gowalla⁷.

Najzákladnejšími prvkami odporúčacích systémov sú používatelia a položky, pričom v prípade LBSN sú položkami tzv. body záujmu (ang. *Points-of-Interest*, skrátene POI). Kontextuálne odporúčanie POI je zaujímavé z dôvodu, že tieto portály sa snažia motivovať používateľov navštevovať nové lokality a tieto skúsenosti následne zdieľať, takže je dôležité používateľovi navrhnúť čo najvhodnejšie možnosti, ale treba pritom dbať na kontext, ako napríklad dostupnosť lokality, teda geografický prvok, ale aj zohľadnenie temporálnych aspektov sa ukazuje ako prospešné [58, 36]. Používatelia častejšie zájdu do baru v sobotu než v iné dni či lyžiar nenavštívi lyžiarske stredisko uprostred leta.

Kvalitné odporúčanie [36] pre zákazníkov LBSN dokáže zlepšiť nielen ich zážitok, ale aj priniesť prospech v podobne služieb reklamným agentúram, ktoré následne môžu jednoducho zacieliť reklamu na potenciálnych zákazníkov. Odporúčanie je pritom založené na využívaní individuálnych preferencií alebo správania v minulosti.

Spomínaným základným kameňom sú POI, ktoré [29, 55] predstavujú špecifickú bodovú lokalitu (napr. múzeum, reštauráciu), pričom jej vieme ako atribút priradiť pár geografických súradníc (zemepisná dĺžka a šírka). Okrem toho môžu byť pridružené ďalšie vlastnosti, napr. kategória.

Transakcie používateľa nad POI prebiehajú tzv. prihlásením (ang. *check-in*) [64, 55], ktorý zaznamenáva ktorý používateľ navštívil ktorú lokalitu v akom čase. Množina prihlásení používateľa v rámci vymedzeného času, napr. jedného dňa, sa nazýva sekvencia prihlásení [64, 55]. Aktivita pred aktuálnou uvažovanou v sekvencii sa nazýva predchodca (ang. *predecessor*), aktivita po následník (ang. *successor*).

K používateľovi je niekedy priradená aj domáca lokalita [54], ktorá vyjadruje miesto bývania používateľa (alebo jednoducho oblasť, kde navštevuje najviac POI). Hlavnou zložkou jeho profilu však zostáva zoznam jeho aktivít.

⁴https://en.wikipedia.org/wiki/Geosocial_networking

⁵<https://foursquare.com/>

⁶<https://www.yelp.com>

⁷<https://en.wikipedia.org/wiki/Gowalla>

4.1 Typické vlastnosti odporúčania POI

V rámci výskumu odporúčacích systémov v doméne LBSN bolo definovaných niekoľko základných charakteristík, ktoré ovplyvňujú preferencie nahlásení používateľa [63, 60, 62]:

- **geografický vplyv** (napr. používateľ je ochotný cestovať za POI iba do určitej vzdialenosti alebo sa pohybuje len v určitých oblastiach),
- **spoločenský vplyv** (LBSN fungujú ako sociálne siete, takže používatelia, napríklad označení ako priatelia, sa navzájom ovplyvňujú),
- **vplyv temporálnych aspektov** (v rozdielnom čase má používateľ rozdielne preferencie, napr. ráno nepôjde do baru a večer pôjde skôr do kina ako do parku),
- **sekvenčný vplyv** (predchodcovia ovplyvňujú, ktoré POI bude následníkom),
- **vplyv kategórií** (každý používateľ má osobitný vkus, v rámci ktorého preferuje určité kategórie POI, napr. má rád indické reštaurácie ale nie čajovne).

Asi najviac skúmaným aspektom z vymenovaných je **geografický**. Práve dôležitosť tohto vplyvu je jeden z najvýraznejších rozdielov oproti klasickým odporúčacím systémom. Dva podobné POI by mali byť hodnotené rozdielne práve na základe polohy [29, 36, 13], keďže používatelia majú tendenciu navštevovať miesta vo svojej blízkosti. [13] ukázal, že používatelia sa zvyknú prihlasovať v okolí niekoľkých centrálnych bodov, pričom kopírujú Gaussovo rozloženie. Dva body záujmu, ktorých vzdialenosť je menšia ako zvolený prah, nazývame susedné POI [64]. Práve tie najčastejšie vytvárajú sekvencie prihlásení [36]. Otázkou ostáva, ako správne modelovať oblasti výskytu používateľa, keďže napríklad mriežkový systém (rozdeliť oblasť na územia s rovnakou veľkosťou) sa neukazuje ako efektívny z dôvodu, že používateľ sa môže vyskytovať na hranici území, rovnako veľkosť jeho preferovanej pod-oblasti sa môže líšiť. Iným prístupom je vytvorenie funkčných oblastí (napr. miesto s veľkým počtom obchodov je nákupné centrum, miesto s veľkým výskytom historických pamiatok turistickým miestom), ale aj tie je potrebné personalizovať pre každého používateľa zvlášť [58]. Taktiež je dôležité poznamenať, že rôzni používatelia môžu mať rôzne záujmy v rovnakej oblasti.

Ďalším aspektom je **spoločenský vplyv**. Niektoré štúdie však tvrdia, že nie je až taký vysoký, ako sa očakávalo [13]. V systéme Gowalla sa len 9,6 % nahlásení priateľov zhodovalo - to naznačuje, že vzájomný vplyv síce existuje, ale je veľmi limitovaný. Pri korelácii používateľa a jeho priateľa sa môže uvažovať sociálna frekvenčná agregácia, rozdelenie odhadu sociálnej frekvencie a výpočet skóre významu [28]. Niektorí používatelia majú však priateľov z iných oblastí či dokonca krajín, čo ešte znižuje ich význam [39].

Temporálne aspekty sa dostali do popredia neskôr, ale ukazujú sa ako významný faktor [58, 36, 18, 60], pričom sa často používajú v kombinácii s kolaboratívnym filtrovaním, keďže niektorí používatelia sa na seba zvyknú svojimi časovými preferenciami podobáť, napr. viacerí si cez pracovné dni ráno kúpia kávu a v sobotu večer idú do baru.

Avšak, temporálne vzory je mnohokrát problém zachytiť kvôli riedkosti dát, teda nízkej frekvencii prihlásení [13, 36, 62, 29, 55]. Spôsobené je to najmä tým, že na rozdiel od systémov, ktoré sa spoliehajú na dáta z WiFi či GPS, systém nahlásení stojí plne na dobrovoľnosti prihlásení používateľa, teda na explicitnej spätnej väzbe. Dôležité je poznamenať, že výrazne menšia hustota dát je vo vidieckych oblastiach oproti tým mestským.

Otvoreným problémom je aj diverzifikácia výsledkov odporúčania [23] - nie len kategorická, ale aj geografická. Klasická zmena zoradenia (hotové zoradenie odporúčaných položiek sa preradí tak, aby zodpovedalo danej diverzifikácii) nie je priamo aplikovateľná na odporúčanie lokality.

4.2 Aktuálne prístupy kontextuálneho odporúčania v oblasti LBSN

Ako bolo načrtnuté v Kapitole 4.1, odporúčacie systémy zameriavajúce sa na POI disponujú niekoľkými špecifickými vlastnosťami, čo znamená aj nové výzvy. Prehľad aktuálnych výskumných prác na túto tému obsahuje Tabuľka 1.

Tabuľka 1: Prehľad aktuálnych riešení POI odporúčacích systémov.

Autor	Názov	Dataset	Metóda	Metrika	Zdroj
Chen Cheng et al.	Fused Matrix Factorization with Geographical and Social Influence in Location-Based Social Networks	Gowalla	MF s predfiltrovaním - geografický (pravdepodobnosť prihlásenia cez <i>Multi-center Gaussian Model</i>) a sociálny vplyv	Presnosť@k, Úplnosť@k	[13]
Gao, Huiji et al.	Exploring temporal effects for location recommendation on location-based social networks	Foursquare	<i>Low-rank</i> MF založená na temporálnych aspektoch, vzdialenosť na základe kosínusovej podobnosti	Presnosť@k, Úplnosť@k	[18]
Shenglin Zhao et al.	Geo-Teaser: Geo-Temporal Sequential Embedding Rank for Point-of-interest Recommendation	Foursquare, Gowalla	Vnorenie temporálnych aspektov - prihlásenia v sekvencií, cyklické vzory pre rôzne dni	Presnosť@k, Úplnosť@k	[64]
Oppokhonov, Shokirkhon et al.	Current location-based next POI recommendation	Foursquare, Gowalla	Kolaboratívne filtrovanie, Jaccard/Person/kosínusova podobnosťou a aktuálna lokalita tvoria skóre	Presnosť, Úplnosť	[39]
Wang, Weiqing et al.	ST-SAGE: A Spatial-Temporal Sparse Additive Generative Model for Spatial Item Recommendation	Foursquare, Twitter	Zohľadňovanie záujmov používateľa a všeobecne používateľov v danej oblasti, skúma spoločný výskyt/obsah položiek	Správnosť@k	[54]
Khanthaapha, Passakorn et al.	Topic-based User Profile Model for POI Recommendations	Foursquare	Vzťah prihlásenia/text LDA (spracovanie používateľových preferovaných kategórií), vyvodenie tém, kosínusova vzdialenosť	Presnosť@k, Úplnosť@k, F-skóre	[29]
Li, Xin et al.	A Time-Aware Personalized Point-of-Interest Recommendation via High-Order Tensor Factorization	Foursquare	Párová interakcia TF (predikovanie kategórie, zoradenie lokalít - bipartitný graf), <i>fourth-order TF</i> (integrácia krátkodobého a dlhodobého temporálneho vplyvu)	Presnosť@k	[36]
Laß, Christopher et al.	Context-Aware Tourist Trip Recommendations	Foursquare	Faktor C (ohľad na kontext) = jednoduchý vážený aritmetický priemer	Porovnanie s dotazníkmi	[35]

Zhang, Jia-Dong a Chow, Chi-Yin	Exploiting Geographical, Social and Categorical Correlations for Point-of-Interest Recommendations	Foursquare, Yelp	Vážená popularita kategórií, odhad rozloženie popularity kategórií, výpočet skóre ako relevantná je kategória	Presnosť, Úplnosť	[62]
Peña, Francisco J.	Unsupervised Context-Driven Recommendations Based On User Reviews	Yelp	Založené na recenziách, opakované modelovanie tém pomocou LDA, <i>non-negative</i> MF	Vlastné skóre	[42]
Liu, Yanchi et al.	Unified Point-of-Interest Recommendation with Temporal Interval Assessment	Foursquare, Gowalla	Správanie záujmov používateľa a jeho vyvíjajúcich sa sekvenciách preferencií s ohodnotenými temporálnymi intervalmi	F-skóre, NDCG	[37]
Wang, Weiqing et al.	TPM: A Temporal Personalized Model for Spatial Item Recommendation	Foursquare, Twitter	Latentný pravdepodobnostný generatívny model tried: latentná premenná <i>topic-region</i> ; vplyv predchodcov, temporálne cykly	Hits@K	[55]
D. Yang et al.	Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs	Foursquare, Gowalla	<i>Non-negative</i> tenzorová faktorizácia (temporálne aspekty)	Správnosť@k, Presnosť, Pokrytie, F1-skóre a i.	[58]
Griesner, Jean-Benoît et al.	POI Recommendation: Towards Fused Matrix Factorization with Geographical and Temporal Influence	Gowalla	MF s temporálnymi a geografickými prvkami	Presnosť@k, Úplnosť@k	[22]
Han, Jungkyu et al.	Geographical Diversification in POI Recommendation: Toward Improved Coverage on Interested Areas	Yelp	Zameriavajú sa na diverzifikácia výsledkov odporúčania POI	Úplnosť@k, ILD EPC PRg PR _{g-r} -@N	[23]
Yuan, Quan et al.	Time-aware Point-of-interest Recommendation	Foursquare, Gowalla	Kolaboratívne filtrovanie	Presnosť@k, Úplnosť@k	[60]
Zhao, Shenglin	Point-of-interest Recommendation in Location-based Social Networks	Foursquare, Yelp	Model sociálneho, kategorického, geografického, sekvenciálneho a temporálneho vplyvu		[63]

Vo všeobecnosti, medzi špecifické a opakujúce sa témy patrí voľba druhu transakcií, ich následné skúmanie z pohľadu analýzy textu, geografických a temporálnych aspektov. Medzi časté metódy patrí kolaboratívne filtrovanie a faktorizácia matíc, pričom zefektívnenie výsledkov je možné dosiahnuť aj ich diverzifikáciou. Spôsoby vyhodnocovania sa obmedzujú takmer výlučne na pár metrík, častokrát však výsledky nepodliehajú celému datasetu, ale len podvýberu z uvedených zdrojov dát.

Druhy transakcií

Väčšina z prác chápala ako transakciu prihlásenie používateľa na určitom POI [13, 18, 64, 39, 54, 29, 36, 35, 37, 58, 22, 60], pričom ako dataset slúžili prevažne dáta z portálov Foursquare či zaniknutého Gowalla. Čím viac mal používateľ prihlásení na určitom POI, tým viac sa jeho hodnotenie lokality považovalo za pozitívne. Niekoľkokrát sa využil aj dataset portálu Yelp [23, 62], pri ktorom sa miesto prihlásení, ktorých záznamy prakticky neexistujú, bralo v úvahu priamo hodnotenie recenzie používateľa (chápalo sa ako ekvivalent k počtu prihlásení [62]). Všetky spomínané práce sa zaoberajú existenciu samotného prihlásenia alebo hodnotenia, takmer vôbec samotným predikovaním hodnoty hodnotenia.

Obohatenie o obsah a analýzu textu

Výnimočne boli tieto dáta obohatené o analýzu textu, zdrojom je najčastejšie Twitter. Zaujímavých, i keď zriedkavejším je Yelp, ktorý bol napríklad použitý v kombinácii s metódou LDA (ang. celým názvom *Latent Dirichlet Allocation*) [42] na modelovanie tém. Recenzie sú pritom reprezentované ako zmes latentných tém - tie sú charakterizované distribúciou nad slovami. Problémom však je, že každé spustenie tohto algoritmu môže produkovať iné výsledky.

Spojenie temporálnych a geografických aspektov

Najčastejšie zapracované vplyvy kontextuálnych faktorov boli geografické a temporálne. Niektoré práce, aby vedeli jednoznačne vyjadriť, aký podiel na zlepšení výsledkov malo zapracovanie temporálnych či geografických aspektov, vypracovali odporúčenie na ich základe oddelene a až následne ich prepojili do jedného systému [58].

Geografické aspekty

Pri geografických aspektoch sa momentálne využíva jednoduchý mriežkový systém minimálne (priestor rozdelený na oblasti v podobe rovnako veľkých štvorcov), miesto toho sú populárne tzv. osobné funkčné oblasti (ang. *Personal Functional Regions*) [58], ktoré predstavujú oblasť vytvorenú centrom a priestor v rámci určitého rádiusu okolo neho

(napr. 0,5 km [23]), kde bola zaznamenaná väčšina používateľových aktivít. Následne sa vyhodnotia preferencie aktivít používateľa v tejto oblasti a ako výsledok máme teda osobné regióny pre používateľa, z ktorých má každý pripísané silné preferencie voči istým kategóriám položiek. Preferencie nenavštívených lokalít sa dajú predvídať pomocou interpolácie. Samotný algoritmus funguje ako pažravé zhľukovanie (ang. *greedy clustering*), ktoré ako centrum označí najnavštevovanejšie POI používateľa a rozšíri región o ďalšie POI na základe GPS súradníc, pričom mu na základe kategórií priradí funkciu a vyhodnotí, či sa jedná o osobný funkčný región na základe zvoleného prahu (ang. *threshold*).

Temporálne aspekty

Uvedené práce ukázali jednoznačný význam temporálnych aspektov v odporúčaní POI. Jedným z najčastejšie sledovaných vplyvov bola zmena preferencií používateľa v čase, teda faktor rozpadu, kedy sa skorším hodnoteniam dáva menšia váha ako novším [60], napríklad používateľ, ktorý mal v minulosti rád čínske reštaurácie, zrazu zmenil preferencie na suši. Celkovo, modelovanie používateľa na základe preferencií kategórií je obľúbený prístup, napríklad [63] ráta skóre kategórií na základe skreslenia (ang. *bias*) kategórie pre používateľa a frekvencie používateľových návštev v POI tejto kategórie. Objavilo sa aj skúmanie cyklických temporálnych vplyvov, teda sezónnosti, kde boli prístupy k zvoleniu intervalov cyklov a ich intervalov rôzne, napr. rozdelenie dňa na hodiny [37, 60], týždňa na hodiny [58], rok na mesiace [37] (tí si všimli pri podnikaní podľa hodín Gaussovské rozdelenie, pri používateľoch Poissonovo rozdelenie). [55] skúsili dva prístupy - zakódovanie cyklických vzorov do binárneho kódu pomocou časového indexovania (veľký problém s nedostatočnou hustotou dát) a rozdelenie času na základe každého cyklického vzoru zvlášť (následne skúma všetky vzory v spojenom aditívnom modeli).

Kolaboratívne filtrovanie

Obľúbená metóda odporúčania je kolaboratívne filtrovanie, pričom ako obsah matice Používatelia x Položky sa používa frekvencia návštev používateľa pre danú položku (počet prihlásení v danej lokalite), prípadne, ak sa miesto prihlásení využívajú ako transakcie recenzie, hodnotenie danej položky používateľom [62]. Samozrejme, väčšina vstupov v tejto matici je prázdna, keďže používateľ dokáže navštíviť len obmedzený počet položiek - to je práve skôr spomínaný problém nedostatočnej hustoty matice.

Na rátanie susednosti sa používajú rôzne susednosti, kosínusova s miernou prevahou [18, 39, 29]. Niektoré prístupy prišli s vlastnou metódou rátania blízkosti [60], kedy sa blízkosť používateľov ráta na základe ich správania počas celej doby zbierania dát, s väčším dôrazom ak sa na rovnakých miestach prihlásili približne v rovnakom čase, takže rozšírili kosínusovu podobnosť zo Vzorec 4 na Vzorec 5, kde CS je kosínusova podob-

nosť, S pozmenená CS , A a B sú vektorom hodnotení položiek používateľom, pričom n je celkový počet položiek vo vektoroch, T počet časových intervalov a L lokalít. Takýto prístup okrem kosínusovej vzdialenosti vyskúšali aj v kombinácii s inými, napr. Pearsonovou vzdialenosťou, ale výsledky boli podobné. Pre uhladenie podobnosti brali v úvahu prihlásenia aj z iných ako aktuálneho intervalu, ale s menšou váhou.

$$CS = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (4)$$

$$S = \frac{\sum_{t=1}^T \sum_{l=1}^L A_{t,l} B_{t,l}}{\sqrt{\sum_{t=1}^T \sum_{l=1}^L A_{t,l}^2} \sqrt{\sum_{t=1}^T \sum_{l=1}^L B_{t,l}^2}} \quad (5)$$

Faktorizácia matíc

Kvôli nedostatočnej hustote dát sú obľúbené metódy na odhalenie latentných vlastností, ktoré pomáhajú zredukovať negatívne následky riedkosti redukciiu priestoru predikcií. [55] zaviedol novú latentnú premennú *topic-region*, [58] používa latentnú koreláciu - tenzor (objekt, zovšeobecnením pojmu vektor, na rozdiel od vektora je možné jeho zložky označiť nielen jedným indexom, ale môže ich mať viacero) troch smerov na mapovanie temporálnych aktivít (týždenný vzor, hodinové intervaly). Používaná je aj faktorizácia matíc, ktorá je *state-of-the-art* metódou v oblasti odporúčania [13, 18, 42, 22]. Príklad využitia je faktorizácia matice (používatelia x POI [22]), obsahujúcej kontextovo pred-filtrované dáta. Jej častým problémom býva náročná škálovateľnosť - riešením môže byť rozdelenie matice pred aplikovaním MF. Rozdiel medzi MF a tenzorovou faktorizáciou (skrátene TF) je, že TF uvažuje kontext ako dodatočnú dimenziu [58].

Niektoré práce ale používajú oveľa triviálnejšie odporúčania, napríklad [39] ráta pravdepodobnosť navštívenia POI ako sumu súčinov podobnosti s iným používateľom a hodnotenie lokality týmto iným používateľom, pričom táto suma je delená sumou všetkých podobností s inými používateľmi.

Diverzifikácia výsledkov

Zaujímavým riešeným problémom je aj diverzifikácia výsledkov. [23] sa zameriaval na proporčnú diverzifikáciu odporúčaní na základe oblastí, pričom rozdelenie medzi oblasti je zvolené na základe frekvencie používateľa v nich. Podobným spôsobom sa dá samozrejme diverzifikovať aj na základe témy, teda kategórie. Zvlášť výhodné na ich riešení je, že sa dá zakomponovať aj do už zhotovených odporúčacích systémov.

Spracovanie a výber dát

Mnohé práce [13] za účelom skúmania samotných kontextuálnych aspektov bez ďalšej záťaže odstránili používateľov, ktorí nedosahovali istý počet prihlásení (napr. 10 [13], 20 [23], 50 [22]), či dokonca podniky, ktoré nemali určitý zvolený prah návštevnosti (5 [23], 20 [13]) alebo len v určitej oblasti (napr. Francúzsko [22], mestá s najväčším počtom prihlásení v datasete [23]).

Čo sa týka deľby dát na tréningovú a testovaciu množinu, najčastejším prístupom bol pomer 80:20 % [55, 37], pričom sa nepredelili takto transakcie ako celok, ale množina pre každého používateľa samostatne [55].

Vyhodnotenie

Na vyhodnotenie výsledkov sa pritom využívali najčastejšie metriky presnosť@k a úplnosť@k [13, 18, 64, 39, 58], ktoré predstavujú presnosť/pokrytie vrámci prvých k odporučených položiek. Postup výpočtu presnosti je Vzorec 6 [36], kde p je presnosť, I_U relevantné položky, I_R odporučené položky - ale v prípade $p@k$ sa berie ako I_R len prvých k odporučení.

$$p = \frac{|I_U \cap I_R|}{|I_R|} \quad (6)$$

Úplnosť@k je možné vypočítať podľa Vzorca 7, kde je značenie rovnaké, okrem toho, že pokrytie je označené r . Taktiež, v prípade pokrytie@k sa pre výpočet berie v úvahu len prvých k odporučených položiek.

$$r = \frac{|I_U \cap I_R|}{|U_R|} \quad (7)$$

4.3 Analýza dostupných datasetov v doméne LBSN

Ako ukázala posledná podkapitola, v doméne kontextového odporúčania LBSN sa využíva niekoľko dostupných datasetov. Tu je základný prehľad tých zaujímavejších a použiteľnejších, ktorých portály sú ešte stále funkčné. Niektoré sú viac či menej úplné a časť potrebuje doložiť aj dáta stiahnuté z webového rozhrania. Datasets rozdelené podľa portálov:

- **Yelp**⁸

- chýba prepojenie používateľov a biznisov cez prihlásenia, tie sú agregované spolu podľa miesta výskytu,
- obsahujú rôzne meta dáta ohľadom podniku,

⁸<https://www.yelp.com/dataset/challenge>

- počet záznamov okolo 146 tisíc,
- zoznam hodnotení ale obsahuje aj ID biznisu, používateľa aj časovú pečiatku,
- rovnako zoznamy tipov,
- celkom veľa informácií o samotných používateľoch.

- **4square⁹**

- NYC Restaurant Rich Dataset (Check-ins, Tips, Tags)
 - * obsahuje menej záznamov, ale aj so slovným hodnotením,
 - * neobsahuje časové pečiatky,
 - * záznamy 3112 používateľov, 3239 miest a 27149 označení na miestach - s 10377 komentármi.
- Global-scale Check-in Dataset
 - * obsahuje viac záznamov, ale nie sú označené hodnotením,
 - * dáta za 18 mesiacov,
 - * 33 278 683 označení, 266 tisíc používateľov, 415 miesta (každé niekoľko tisíc označení), 77 krajín,
 - * označenia obsahujú aj časovú pečiatku a ID používateľa.

- **TripAdvisor**

- dostupné datasety sú len malej veľkosti¹⁰¹¹, avšak je tu potencionálna možnosť sťahovať dáta prostredníctvom webového rozhrania,
- dostupné datasety obsahujú atribúty ako jednotlivé zložky hodnotenia, časovú pečiatku hodnotenia, slovné komentáre, rôzne meta informácie o položke,
- jedná sa vlastne o záznamy hodnotení návštev jednotlivých atrakcií v rôznych mestách, podľa ktorých sú rozdelené.

⁹<https://sites.google.com/site/yangdingqi/home/foursquare-dataset>

¹⁰<http://times.cs.uiuc.edu/~wang296/Data/>

¹¹https://www.researchgate.net/publication/308968574_TripAdvisor_Dataset

4.4 Diskusia

V Kapitole 4 sme preskúmali charakteristiku domény odporúčania POI - jej aktuálne problémy a návrhy ich riešenia, ktorá sa vyznačuje širokým uplatnením kontextuálnych faktorov. Zakomponovanie geografických faktorov, najmä tzv. osobných funkčných regiónov, výsledky odporúčania výrazne zlepšuje a momentálny výskum zanecháva dojem, že s temporálnymi aspektami je to rovnako. Najmä sezónnosť a faktor rozpadu sa ukázali ako smerodajné. Avšak, na rozdiel od geografických aspektov sa im venuje oveľa menej priestoru vo výskume a to poskytuje možnosti pre hľadanie nových prístupov.

- Cyklickosť je pri temporálnych aspektoch populárnou témou, ale uvažuje sa len globálne, v kontexte závislosti od geografickej oblasti sú z väčšej časti zanedbané, čo je zaujímavá výzva, keďže mnoho zo spomínaných portálov využívajú používatelia z celého sveta, kde klíma a kultúrne zvyky rozhodne nie sú jednoliate.
- Zatiaľ čo mnoho prác sa zaoberá postupnými zmenami preferencií jednotlivých používateľov, všeobecná zmena trendov v spoločnosti, teda stúpanie a klesanie popularity istých kategórií položiek či konkrétnych položiek, býva zanedbávaná. Už vôbec nie je uvažovaná v kontexte lokálnych trendov len v určitých uvažovaných oblastiach, hoci napríklad v rôznych krajinách sa aktuálne trendy môžu rôzniť.

Pri doterajších riešených problémoch sa ukázali ako stále populárne metódy kolaboratívneho filtrovania (najčastejším spôsobom rátania vzdialeností používateľov bola kosínusová vzdialenosť) a rôzne latentné metódy, ako tenzorová a maticová faktorizácia.

Tieto prístupy však majú svoje nedostatky, jedným z nich je, že väčšina prác uvažovala prihlásenie používateľa ako transakciu, tie sú však pritom veľmi nespoľahlivé, keďže pravidelné prihlasovania nútia používateľa podávať častú explicitnú spätnú väzbu, čo mnohí nie sú ochotní podstúpiť. Preto si myslíme, že by mohlo byť zaujímavé použiť ako vstupy pre používateľa záznamy vyžadujúce menšiu mieru interakcie, napríklad recenzie, ktoré sú zadávané jednorazovo. Okrem toho však recenzie umožňujú aj posunutie úlohy od predikcie pravidelných opakujúcich sa návštev k predikcii návštevy zvyčajne pre používateľa nového bodu záujmu. Dojem z POI reflektuje v kvantitatívnom alebo opisnom hodnotení.

Štandardne sa využívajú datasety z troch internetových portálov - Foursquare, Gowalla a Yelp, pričom prvé dva považujú za primárne transakcie prihlásenia, len posledný recenzie s hodnotením. Pri odporúčaní bodov záujmu sa iba zriedkavo vyskytuje úloha predikcie samotnej hodnoty hodnotenia, zvyčajne ide iba o predikciu návštevy.

5 Návrh metód odporúčania na základe sezónnosti a dlhodobých trendov špecifických pre lokalitu

Cieľ našej práce je vytvorenie metód, ktoré zohľadnia časové aspekty. Ako bolo v Kapitole 3 vysvetlené, tento aspekt býva často zanedbávaný a podceňovaný. V Kapitole 3 a 4 sme preskúmali aktuálne riešenia a otvorené problémy v odporúčaní POI na základe cyklických vzorov a faktorov rozpadu.

Ako jeden z problémov sme uviedli, že prihlásenie ako vstup od používateľa kladie naňho vysoký tlak v podobe vyžadovania častej explicitnej spätnej väzby, ktorému len máloktoľ použivateľ vyhovie, a preto kompletne záznamy častokrát chýbajú a nie sú stopercentne spoľahlivé [13, 36, 62, 29, 55]. Preto na rozdiel od väčšiny existujúcich prístupov si za svoju úlohu nekladíme predikovať najbližšie prihlásenie na používateľovi už známom POI, ale iný druh explicitnej spätnej väzby, ktorou je pridanie recenzie k danému POI. Takáto spätná väzba je jednorazová a teda menej náročná na získanie zo strany používateľa. Predpokladáme, že pridanie recenzie vo väčšine prípadov zodpovedá prvej návšteve používateľa v novom podniku, preto môžeme túto úlohu formulovať ako odporúčanie nového bodu záujmu. Samozrejme, aj táto voľba má svoje limity, napríklad nevieme stopercentne, či používateľ píše recenziu približne v čase, keď podnik navštívil, alebo s veľkým odstupom. Na druhej strane, aj v prípade, že hodnotí podnik s odstupom, znamená to, že jeho hodnotenie je stále aktuálne (napr. bol v reštaurácii pred rokom a teraz napísal pozitívnu recenziu - dá sa predpokladať, že jej návštevu aj momentálne hodnotí ako pozitívnu skúsenosť).

Našou hlavnou výskumnou otázkou je pritom, akým spôsobom vieme zohľadniť temporálne vplyvy na základe geografickej oblasti pri odporúčaní prvej návštevy POI. V tejto oblasti chceme pritom bližšie preskúmať dve oblasti:

- H1: Vymedzenie vzorov sezónnosti pre jednotlivé kategórie POI podlieha závislosti od geografickej oblasti a zvýši presnosť a pokrytie pri odporúčaní nových bodov záujmu, vyjadrených formou pridania novej recenzie.
- H2: Kategórie POI podliehajú dlhodobým trendom v spoločnosti, teda sú závislé aj od geografickej oblasti a ich využitie zvýši presnosť a pokrytie pri odporúčaní nových bodov záujmu, vyjadrených formou pridania novej recenzie.

Prvá hypotéza sa zaoberá sezónnosťou, teda cyklickými vplyvmi, ale nie na úrovni používateľov, ale kategórií podnikov v jednotlivých geografických oblastiach. Je logické, že popularita zmrzlinární je iná cez leto a iná cez zimu, ale netreba zabúdať na to, že leto a zima sú iné v Las Vegas a iné v Anchorage na Alijaške.

Druhá hypotéza sa odkazuje na prípady, keď istá činnosť viažuca sa so špecifickou kategóriou aktivít je populárna v spoločnosti, ale tento trend je len dočasný. Ako príklad si môžeme zobrať predajňu CD nosičov s hudbou - kým bolo toto médium nové a inovatívne, postupne výskyt zákazníkov v predajni stúpal, ale časom sa trendy v nákupoch v hudobnom priemysle zmenili, presunuli do online sveta, a tak prirodzene aj počet zákazníkov začal klesať. Avšak, v rôznych oblastiach sa môžu tieto trendy líšiť.

Pre zodpovedanie hypotéz sme navrhli metódu odporúčania, ktorá:

- zohľadňuje primárne temporálny kontext,
- zohľadňuje sekundárne geografický kontext,
- kombinuje kolaboratívne a obsahové odporúčanie,
- vychádza zo *state-of-the-art* metódy faktorizácie matíc,
- pokladá recenzie z oblasti LBSN za transakcie,
- odporúča POI.

Najjednoduchšie zavedenie temporálnej špecifickosti podľa lokality by sa riešilo natrénovaním samostatných modelov pre jednotlivé lokality. Avšak, to by prinieslo nevýhody napríklad v podobe rozdelenia používateľov na viacero inštancií (ak sa ich transakcie vyskytujú vo viacerých lokalitách) a zníženie objemu dát pre modelovanie podobností. Preto využívame explicitné zavedenie temporálnych aspektov lokality do odporúčania, tzv. jeden model pre viaceré lokality.

5.1 Konceptuálny návrh metódy

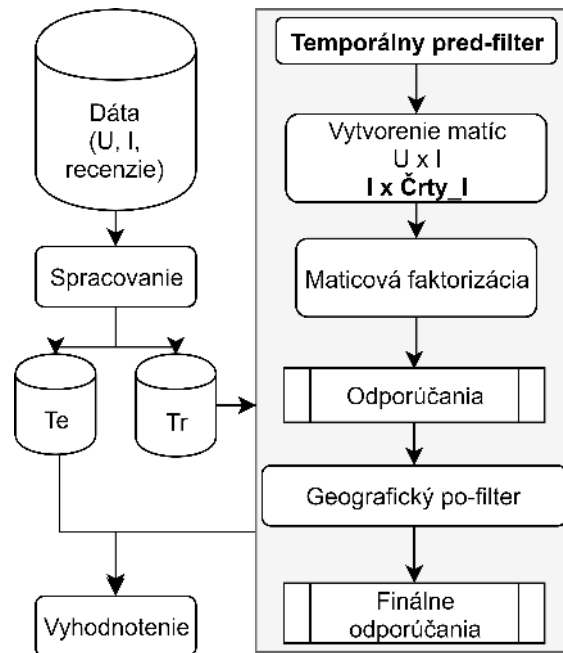
Pre overenie každej z dvoch hypotéz sme použili dva rôzne prístupy. Tie sú nadstavbou nad základnú metódu odporúčania v podobe kolaboratívnej maticovej faktorizácie [27]:

- prístup založený na modelovaní,
- prístup založený na po-filtrovaní.

S tým, že oba prístupy zahŕňajú rozdelenie dát na trénovacie a testovacie, finálne po-filtrovanie na základe lokality a v prípade vplyvu sezónnosti aj temporálny pred-filter na základe kalendárneho mesiaca. Takýto pred-filter sa označuje ako reprezentatívny, heuristický s kategorickou premennou (zvažovaný interval).

Je dôležité poznamenať, že dáta pre trénovanie a testovanie sú kvôli špecifickému typu odporúčania rozdelené na základe časovej pečiatky transakcie - posledný rok dát (približne 20 %) je vyhradený na testovanie.

Prehľad navrhutej metódy na základe modelovania je zobrazený na Obrázku 6. Základ prístupu je na vytvorení matice črt jednotlivých položiek, ktorá sa používa v maticovej faktorizácii ako ďalší vstup pri modelovaní. Črty (založené na kategóriách položky) sú jedinečné pre každú metódu, zhodujú sa ale vo využití črty pre označenie geografickej oblasti.



Obr. 6: Konceptuálny návrh metódy založenej na modelovaní (Tr = trénovacia množina, Te = testovacia množina).

Prehľad navrhutej metódy na základe po-filtrovaní je zobrazený na Obrázku 7. Na rozdiel od predchádzajúceho prístupu, črty položiek nevstupujú do modelovania ako matica. Vplyv kategórií položiek sa uplatňuje až po prvotnom vygenerovaní odporúčaní pomocou maticovej faktorizácie. Tieto odporúčania sa na základe črt z kategórií položiek preusporiadajú.

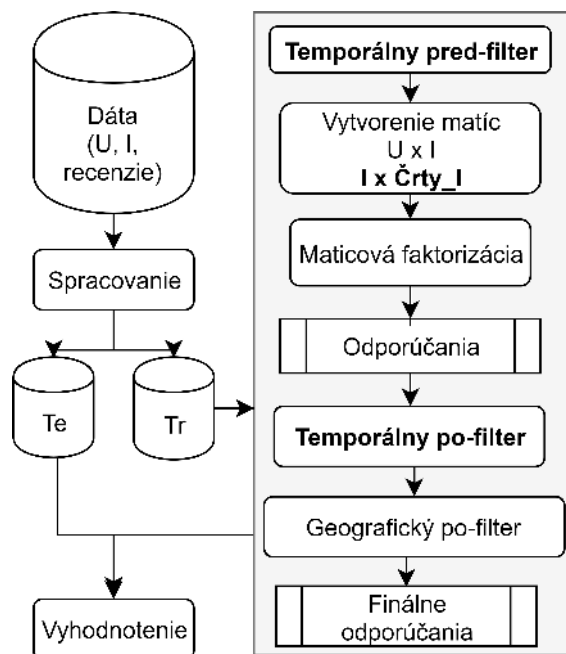
Z odporúčaní sú vždy odstránené už navštívené POI, keďže odporúčame prvú návštevu.

Na záver sa vykoná spomínaný po-filter na základe lokality - odporúčania v blízkosti bodov záujmu navštívených používateľom v minulosti sú uprednostnené.

5.2 Výber a spracovanie dát

Pracujeme s recenziami (hodnoteniami od 1 do 5), lebo je pravdepodobnejšie, že sa používateľ odhodlá podnik ohodnotiť raz, ako že sa bude vždy pri jeho návšteve prihlasovať v aplikácii.

Záznamy bez kategórií sú odstránené, rovnako ako podniky bez záznamov, podobne



Obr. 7: Konceptuálny návrh metódy založenej na po-filtrovaní (Tr = trénovacia množina, Te = testovacia množina).

ako v práci [29]. Berieme v úvahu len používateľov s 5 a viac hodnoteniami. Podobné riešenie zvolili aj autori v práci [23]. Opakované recenzie jedného používateľa na jeden podnik sú vylúčené (minimálna strata v jednotkách kusov), pričom sa zachovávajú tie novšie.

Vytvoríme dve sady dát pre optimalizáciu a validáciu riešenia:

- oblasť bez výrazných sezónnych vplyvov počasia (A),
- oblasť s výraznými sezónnymi vplyvmi počasia (B).

Vytvoríme ďalšie tri sady pre overenie:

- dve oblasti bez výrazných sezónnych vplyvov počasia (Z a W),
- dve oblasti s výraznými sezónnymi vplyvmi počasia (X a Y),
- ich kombinácia (XYZW).

Pod oblasťou si môžeme predstaviť napr. štát.

Ako testovacia množina slúži posledný rok použitých záznamov. Pre testovanie Hypotézy 2 o dlhodobých trendoch kategórií je použitá ako celok, ale pri testovaní sezónnosti (Hypotéza 1) je rozdelená na 12 častí podľa mesiaca, resp. vznikne z nej dvanásť testovacích množín a k nim vznikne aj 12 trénovacích množín, vždy kombináciou pôvodnej

trénovacej množiny a testovacích množín predchádzajúcich mesiacov. Napríklad trénovacia množina pre mesiac február bude pozostávať z pôvodnej trénovacej množiny (dáta pred aktuálnym rokom) spolu s testovacou množinou január (predchádzajúca testovacia podmnožina). Konečné výsledky sa priemerujú pre všetkých 12 mesiacov.

5.3 Základné odporúčanie

Základná metóda pre odporúčanie využíva maticovú faktorizáciu na základe kolaboratívneho filtrovania. Do nej sa pri skúmaní hypotéz pridávajú črty alebo sa jej výsledky po-filtrujú, aby sa zohľadnil kontextuálny vplyv. Maticová faktorizácia zároveň slúži na redukciu priestoru predikcií [13, 18, 42, 22].

Kľúčovou je matica $U \times I$, pričom jej hodnoty budú predstavovať hodnotenie používateľa pre dané POI rovnako ako v [62]. Tieto hodnotenia sú pritom za účelom posilnenia modelu normalizované na:

- 1 (pozitívna spätná väzba, vyššie kvantily zvolených hodnotení konkrétneho používateľa),
- 0 (základná hodnota, predstavujúca neexistujúcu interakciu),
- -1 (negatívna skúsenosť, najnižší kvantil hodnotení konkrétneho používateľa).

Stratová funkcia (ang. *loss function*) hrá kľúčovú rolu pri mapovaní hodnôt jednej alebo viacerých premenných na reálne čísla, reprezentujúce "stratu", pričom túto stratu sa snažíme minimalizovať. Bežne používanými stratovými funkciami sú:

- logistická (užitočná pri prítomných aj pozitívnych, aj negatívnych interakciách),
- BPR (Bayesovské personalizované hodnotenie, ang. *Bayesian Personalized Ranking* [45]):
 - vhodný pre implicitnú spätnú väzbu modelov trénovaných cez negatívne vzorky,
 - párová strata,
 - maximalizuje predikčný rozdiel medzi pozitívnym príkladom a náhodným negatívnym príkladom,
 - vhodné v prípade, že sú prítomné len pozitívne interakcie,
- WARP (vážená približná pozícia v poradí, ang. *Weighted Approximate-Rank Pairwise* [56]):
 - možná aj implicitná spätná väzba,

- vo všeobecnosti je jej výkon lepší ako BPR,
 - maximalizuje poradie pozitívnych príkladov opakovaním vzorkovania negatívnych príkladov, pokiaľ nie je nájdené narušenie poradia,
 - vhodné, keď sú prítomné len pozitívne interakcie,
 - vhodné, ak je snaha optimalizovať najvyššie položky v zozname odporúčaní,
 - na rozdiel od BRP, negatívne interakcie nie sú vybrané náhodne, ale spomedzi tých negatívnych položiek, ktoré narušajú žiadané poradie položky, dané stavom modelu,
- k-OS WARP (k-ty rád štatistickej straty, modifikácia WARP, ktorá používa k-tu pozitívnu vzorku pre každého používateľa ako základ na párové aktualizácie [57]).

Keďže, ako je spomenuté v zozname, WARP stratová funkcia je najvhodnejšia pri optimalizácii najvyšších položiek v zozname odporúčaní, pričom podľa analýzy najpoužívanejšími metrikami sú presnosť@k a pokrytie@k, rozhodli sme sa ju použiť pri maticovej faktorizácii.

Čo sa týka učiacich programov (ang. *learning rate schedules*) z možností *adagrad* [15] a *adadelta* [61] sme sa rozhodli pre *adadelta*. Dôvodom je, že zatiaľ čo *adagrad* adaptuje proces učenia na parametre, pracuje na malých zmenách pre parametre často vyskytujúcich sa črt, a väčšie zmeny pre parametre menej častých črt, čo je vhodné pre riedke ráta, *adadelta* ako jeho rozšírenie pridáva aj redukovanie jeho agresivity, monotónne znižujú rýchlosť učenia, obmedzuje úseky nahromadených minulých gradientov na presnú veľkosť¹².

5.4 Temporálne modelovanie

Temporálnym modelovaním LightFM kombinujeme kolaboratívny a obsahový prístup pri odporúčaní. Používateľov a položky modelujeme ako lineárnu kombináciu latentných faktorov ich kontextuálnych črt [34].

Každá položka (používateľ) je reprezentovaná sadou črt (ang. *feature vectors*). Latentná reprezentácia položky (používateľa) je daná sumou latentných vektorov jej črt – *vektor()*. Skreslenie položky (používateľa) je dané sumou skreslení (ang. *bias*) črt – *bias()*.

Pravdepodobnosť používateľovej transakcie nad položkou $\hat{r}$ je definovaná vzťahom 8, kde σ značí použitú funkciu (WARP), u používateľa, i položku, *vektor()* sumu latentných vektorov črt používateľa/položky a *bias()* sumu skreslení črt používateľa/položky.

$$\hat{r}(u, i) := \sigma(\text{vektor}(u) \cdot \text{vektor}(i) + \text{bias}(u) + \text{bias}(i)) \quad (8)$$

¹²<http://runder.io/optimizing-gradient-descent/index.html#adagrad>

Každej položke (používateľovi) sa teda pridá zoznam črt. Každej položke (používateľovi) pridávame aj črtu, ktorá reprezentuje jej identifikátor a tak vytvoríme maticu identít. Tento hybridný model maticovej faktorizácie, vďaka využitiu obsahových črt, umožňuje odporúčaniu lepšie zvládať príchod nových používateľov a položiek pomocou zovšeobecnenia.

Pre každú z hypotéz zostrojíme špecifické črty, pričom sa ich snažíme jednotlivo otestovať a následne nájsť ich najlepšiu kombináciu. Tá vždy obsahuje aspoň jednu inštanciu pre reprezentáciu oblasti a kategórie.

5.4.1 Sezónnosť

Do modelovania sa pri sezónnosti zapracujú nasledovné kontextuálne faktory ako matica črt položky:

- mesiac ako:
 - číslo,
 - cyklická ordinálna hodnota, dve premenné v podobe sínusu a kosínusu,
- štát ako *one-hot* kódovanie (každá kategorická hodnota je reprezentovaná binárnou črtou),
- kategórie ako:
 - *one-hot* kódovanie,
 - ich aritmetický/vážený (väčší dôraz na menej frekventované kategórie) priemer v danej geografickej oblasti (štáte):
 - * pomer priemerného počtu recenzií v danom mesiaci voči priemernému počtu recenzií počas roka,
 - * pomer priemerného počtu recenzií v danom mesiaci, mesiaci pred ním a po ňom (napr. pre január aj december a február) voči priemernému počtu recenzií počas roka.

Pričom pri extrahovaní črt z kategórii POI sa ignorujú kategórie so zanedbateľným priemerným počtom recenzií za rok, teda tie s príliš nízkou frekvenciou výskytu.

Črty sú zvolené tak, aby reprezentovali znaky lokálnej sezónnosti, pričom v našom prípade ako intervaly berieme mesiace. Interval mesiaca sa používa bežne v iných prácach [37]. Preto sú mesiacu venované hneď dve premenné, z ktorých sa vyberie jedna. Cyklická verzia bola navrhnutá z dôvodu, že samotné mesiace v roku sa cyklicky obmieňajú. Cyklická premenná umožňuje zachytenie vzťahu medzi decembrom a januárom

(vzdialenosť december-január je rovnaká ako medzi hocktorými inými susednými mesiacmi). Oblasť reprezentuje zakódovaný štát.

Kategóriu sa snažíme zakódovať aj do numerickej hodnoty, skóre. Účelom je vyjadriť jej popularitu pre jednotlivé mesiace voči všeobecnej popularite v priebehu roka. Takáto hodnota by mala vyjadrovať, aká je pravdepodobná návšteva danej kategórie v aktuálnom období, a či ju teda v aktuálnom období uprednostniť alebo nie.

Aby sme zachytili vývoj položiek v čase, v matici črt položiek budeme vytvárať záznamy jednotlivých položiek pre rôzne časové obdobia (napr. pre položku, ktorej zohľadňujeme sezónnosť po mesiacoch, vytvorili by sme záznamy tejto položky pre rôzne mesiace, kde črty by boli spočítané z aktivít počas týchto mesiacov [34]). Teda vykonáme pred-filtering s dátami z podobných časových úsekov, pričom položku vo viacerých úsekoch chápeme ako viacero jej inštancií. Pre každý POI sa vytvorí 12 inštancií, každá pre jeden mesiac. Transakcie POI sú medzi ne rozdelené, tým sa posilňujú viac frekventované mesiace. Samozrejme, pri vyhodnotení sa všetky inštancie opäť považujú za jednu položku.

5.4.2 Dlhodobé trendy

Do modelovania dlhodobých trendov sa zapracujú nasledovné kontextuálne faktory ako matica črt položky:

- štát (*one-hot* kódovanie),
- kategórie ako:
 - *one-hot* kódovanie,
 - ich aritmetický/vážený (väčší dôraz na menej frekventované kategórie) priemer v danej geografickej oblasti (štáte):
 - * rozdiel počtu recenzií oproti predchádzajúcemu roku (trojica črt pre posledné tri roky),
 - * rozdiel transformovaný na percentuálnu zmenu (trojica črt pre posledné tri roky),
 - * rozdiel normalizovaný na 1 (stúpanie), 0 (bez rozdielu) a -1 (pokles) (trojica črt pre posledné tri roky).

Črty sa zameriavajú na zmeny oproti predchádzajúcej sezóne, pričom za sezónu sa považuje jeden rok (zvyčajne dostatočný počet transakcií a zároveň dostatočne dlhé obdobie na vykreslenie pozorovateľnej zmeny). Sledujeme viacero sezón aby sme vedeli pozorovať dlhodobý trend, nie iba prípadnú nárazovú anomáliu.

Pri extrahovaní črt z kategórii POI sa ignorujú kategórie so zanedbateľným priemerným počtom recenzií za rok, teda tie s príliš nízkou frekvenciou výskytu.

5.5 Temporálny po-filter

Temporálny po-filter nepoužíva modelovanie na základe črt položiek. Zo základného odporúčania (maticová faktorizácia na kolaboratívnych princípoch) sa zoberie niekoľko prvých odporúčaní a tie sa preusporiadajú na základe kontextuálneho skóre.

Ako výsledné poradie sa uvažuje kombinácia temporálneho po-filtru a pôvodného vygenerovaného poradia, ktorá je vyvážená. Tento pomer sa zistí experimentálne pri trénoch (zohľadňuje sa aj pomer 1:1).

5.5.1 Sezónnosť

Pre po-filtrovanie sa využívajú rovnaké numerické črty kategórií, ako pri modelovaní. V tomto prípade sa považujú za získané skóre.

Pričom pri extrahovaní črt z kategórii POI sa zanedbávajú kategórie so zanedbateľným priemerným počtom recenzií za rok, teda tie s príliš nízkou frekvenciou výskytu.

Okrem modelovania sa pri sezónnosti využíva aj pred-filtrovanie podľa mesiacov. Pre každý POI sa vytvorí 12 inštancií, každá pre jeden mesiac. Transakcie POI sú medzi ne rozdelené, čím sa posilňujú viac frekventované mesiace. Samozrejme, pri vyhodnotení sa všetky inštalácie opäť považujú za jednu položku.

5.5.2 Dlhodobé trendy

Používané črty kategórií POI sú rovnaké, ako pri modelovaní. Skóre sa vypočíta na základe vždy trojice črt pre posledné sezóny ako ich pomer 4:2:1 od poslednej k najdávnejšej.

Pričom pri extrahovaní črt z kategórii POI sa zanedbávajú kategórie so zanedbateľným priemerným počtom recenzií za rok, teda tie s príliš nízkou frekvenciou výskytu.

5.6 Geografický po-filter

Zpracovanie geografických aspektov prebieha geografickým po-filtrovaním. Tento krok je potrebný, aby sa používateľovi odporúčali primárne podniky v jeho fyzickom dosahu.

Preto je pre každé odporúčané POI na základe temporálnych aspektov vyrátaná vzdialenosť ku každej z predtým navštívených položiek daným používateľom. Ak aspoň jedna zo vzdialeností je menšia/rovná určenej hraničnej konštante (rádus), bude dané odporúčané POI presunuté na začiatok zoznamu všetkých odporúčaní. Veľkosť rádusu od predtým

navštívených položiek zistíme pomocou experimentálneho nastavovania počas ladenia parametrov na validačnej sade.

Presun geograficky blízkych položiek ja navrhnutý tak, aby bol stabilný pre všetky metódy zaoberajúce sa temporálnymi aspektami a dal dôraz práve na temporálny vplyv. Odporúčané položky na základe temporálnych aspektov sa rozdelia na dva zoznamy - či sa nachádzajú alebo nenachádzajú v danom rádiuse od už navštívených POI. Poradie v nových zoznamoch je zachované podľa pôvodného vygenerovaného temporálnym odporúčaním. Následne sa celý zoznam položiek v rádiuse od navštívených položiek presunie pred zoznam tých vzdialenejších a tak sa vytvorí finálny zoznam odporúčaní.

Pri už navštívených položkách pritom nepozeralme na ich hodnotenia ani podobné parametre, berieme ich jednotne, keďže negatívnym hodnotením používateľ zvyčajne vyjadruje hodnotenie POI, nie oblasti umiestnenia POI ako takej. Napríklad, keď si používateľ chodí kúpiť kávu pred prácou, nebude chodiť o kilometer ďalej len preto, že mu v kaviarni cez ulicu od práce nechutila - pravdepodobne navštívi inú kaviareň v blízkosti tej prvej.

5.7 Vyhodnotenie

Podľa analýzy (Kapitola 4) najpoužívanějšími metrikami v tejto oblasti sú presnosť@k (rovnica 9) a pokrytie@k (rovnica 10), rozhodli sme sa preto použiť ich percentuálnu hodnotu ako našu metriku. Každú z nich aplikujeme na vyhodnotenie prvých k odporúčaní – 1, 3, 5 a 10. Tým zároveň kontrolujeme, či sú relevantné odporúčania v prvých alebo nižšie umiestnených odporúčaníach.

$$p@k = \frac{|I_U \cap I_R|}{\min(k, |I_R|)} * 100 \quad (9)$$

$$r@k = \frac{|I_U \cap I_R|}{|I_R|} * 100 \quad (10)$$

k - počet zohľadnených odporúčaní, $p@k$ - presnosť pre prvých k odporúčaní, $r@k$ - pokrytie pre prvých k odporúčaní, I_U relevantné položky, I_R odporúčené položky.

Vždy sa medzi sebou porovnáva základné odporúčanie a jeho nadstavba (temporálne modelovanie alebo po-filtrovanie). Súčasne sa všetky riešenia porovnávajú s po-filtrovaním na základe popularity podnikov. Priebežne sa vyhodnocuje aj pretrénovanie na základe odporúčaní, ktoré sa už nachádzali v trénovacej množine.

Finálne vyhodnotenie najlepších metód prebieha na troch datasetoch - miestach ná- chylných na temporálne zmeny, miestach s menej očakávaným temporálnym vplyvom a ich kombinácii. Vyhodnocujú sa pritom najlepšie finálne metódy a to porovnaním základnej maticovej faktorizácie, zohľadnením lokálnych temporálnych vplyvov a globálnych temporálnych vplyvov (črty netvoríme na základe geografických oblastí, ale globálne).

5.8 Sumarizácia navrhnutej metódy

V tejto kapitole sme na základe získaných poznatkov navrhli, ako zakomponovať geografické a časové vplyvy do základnej metódy odporúčania v podobe maticovej faktorizácie založenej na kolaboratívnych princípoch. Navrhli sme sériu potencionálnych temporálnych črt a dva prístupy, ako ich použiť - modelovanie a temporálne po-filtrovanie.

Metóda modelovania sa vyznačuje využitím vektorov temporálnych a iných črt položiek ako vstup pre maticovú faktorizáciu. Po-filtrovanie na základe skóre vyrátaného z rovnakých črt mení poradie už vygenerovaných odporúčaní. Pre jeden z otvorených problémov sa používa aj temporálne pred-filtrovanie. Na konečné výsledky sa vždy aplikuje geografický po-filter, aby boli uprednostnené položky vo fyzickej blízkosti používateľa.

Tieto metódy sa používajú na vyhodnotenie hypotéz ktoré hovoria, že zohľadnenie sezónnosti a dlhodobých trendov kategórií zlepšuje výsledky odporúčania. Špecifické sú však najmä v tom, že spomínané sezónnosti a trendy zohľadňujú na lokálnej úrovni, v daných oblastiach resp. štátoch. Preto sú na konci porovnávané nie len s výsledkami základnej metódy, ale aj s náhradou lokálnych trendov v podobe globálnych trendov.

Riziká zvoleného prístupu vidíme najmä v možnom zlom navrhnutí veľkosti intervalov pri skúmaní časovo závislých trendov a málo komplexných črtách. Takisto je problematické správne zvoliť vhodnú dátovú sadu pre overenie.

6 Výber a opis dátovej sady pre overenie navrhnutých metód odporúčania

Pre overenie metódy na personalizované odporúčanie, ktoré berie v úvahu aj kontextuálne vplyvy, je dôležité vybrať vhodný dataset a vykonať jeho analýzu, pričom treba brať ohľad na špecifický druh transakcií a atribútov, ktoré odporúčanie na základe temporálnych aspektov vyžaduje.

6.1 Výber datasetu

Na základe analýzy dostupných datasetov v doméne LBSN v Kapitole 4.3 sme sa rozhodli pre oficiálnu dostupnú množinu dát portálu Yelp¹³ - Yelp Dataset Challenge¹⁴, ktorá je voľne použiteľná na akademické účely pre študentov. Výhodou je aj to, že dataset je pravidelne aktualizovaný už niekoľko rokov. Hlavným dôvodom tejto voľby sú však transakcie v podobe recenzií, keďže naša metóda sa spolieha na tento druh explicitnej spätnej väzby ako na zdroj transakcií.

Samotná aplikácia bola založená v roku 2004 so základnou myšlienkou pomáhať používateľom nájsť pre nich vhodné miesta a služby (kino, doktora, reštauráciu a pod. - čiže POI). Zisky spoločnosti pritom pochádzajú z predaja reklamy lokálnym podnikom. Momentálne má mesačne v priemere 34 miliónov unikátnych používateľov, ktorý dokopy napísali už viac ako 150 miliónov recenzií.

6.2 Opis datasetu

Dostupný dataset pozostáva z niekoľkých JSON súborov:

- POI (podniky),
- používatelia,
- recenzie,
- tipy,
- počty prihlásení v POI,
- fotografie.

¹³<https://www.yelp.com>

¹⁴<https://www.yelp.com/dataset/challenge>

Fotografie pre nás nie sú podstatné, rovnako tipy, z ktorých textu by bolo potrebné dolovať informácie pomocou spracovania prirodzeného jazyka. Taktiež počty prihlásení v jednotlivých POI sa obmedzujú na priemerné hodiny pre hodiny počas týždňa, čo je pre nás irelevantná informácia.

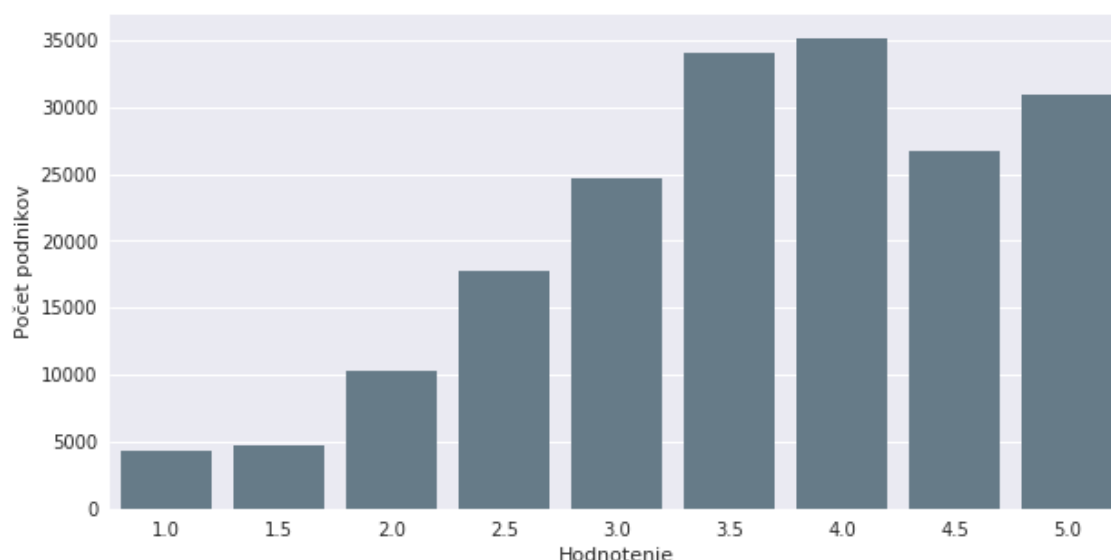
Naopak, najviac podstatné sú informácie o 1 518 169 používateľoch, 188 593 POI a 5 996 996 recenziách, ktoré obsahujú aj číselné hodnotenie.

6.2.1 Používatelia

Počet atribútov pre používateľa je okolo 15. Niekoľko z nich má vysoký počet chýbajúcich hodnôt medzi používateľmi, napríklad priatelia používateľa. Najzaujímavejšie atribúty sú fanúšikovia používateľa a rôzne formy hodnotenia jeho profilu.

6.2.2 Podniky

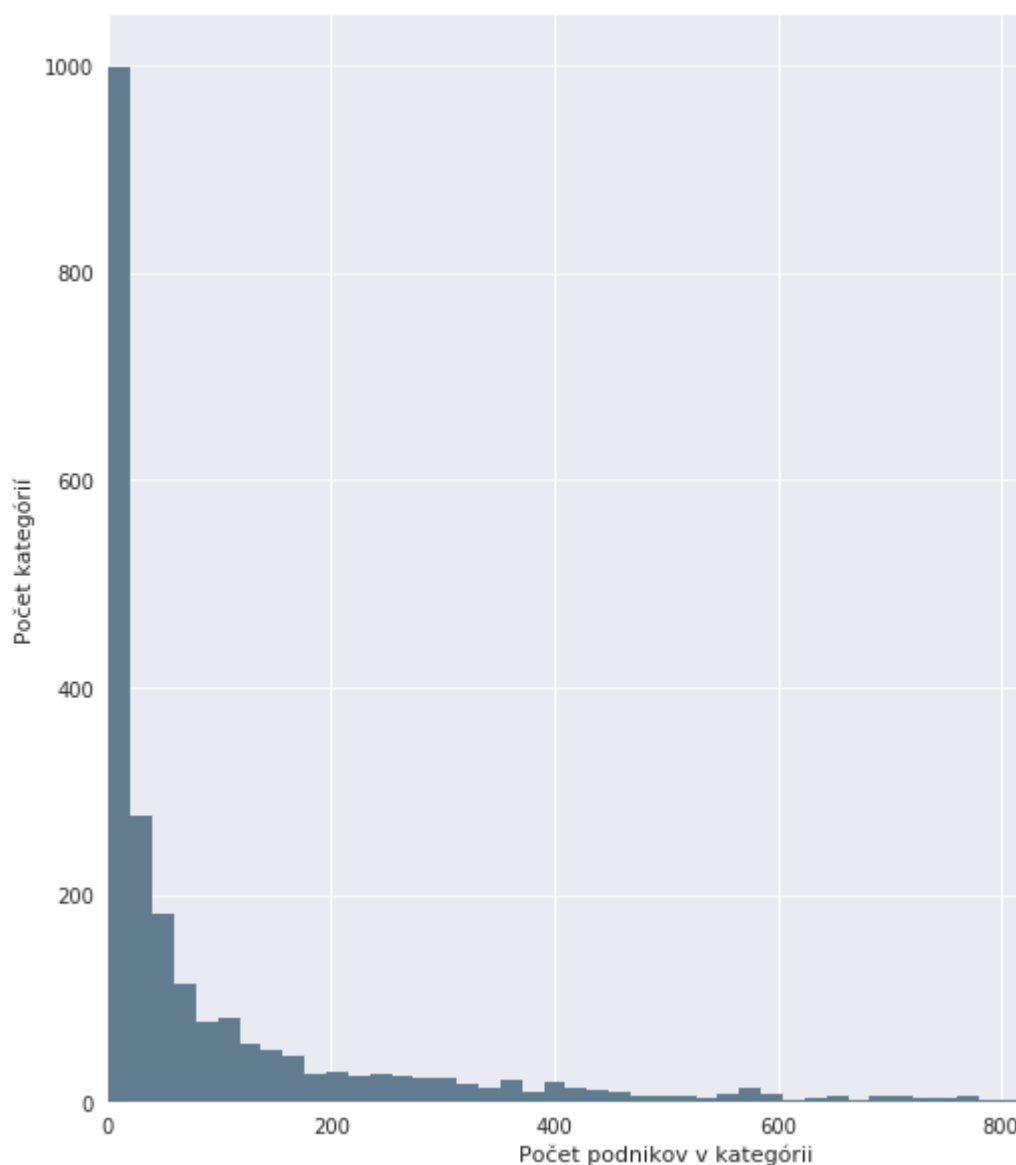
Podniky obsahujú viac ako 50 atribútov, z toho väčšina má však mnoho chýbajúcich hodnôt. Použiteľné metadáta, teda tie, ktorých je dostatok, je informácia, či je podnik ešte v prevádzke, otváracie hodiny pre jednotlivé dni, zemepisná šírka a dĺžka, poštové smerové číslo, adresa, štát, mesto, meno, počet recenzií, hodnotenie (Obrázok 8), niekoľko informácií o službách (možnosť platby kartou, dostupné parkovanie, cenová kategória) a kategórie pod ktoré podnik spadá.



Obr. 8: Počet podnikov podľa hodnotenia.

Najčastejšie kategórie sú reštaurácia, obchod, jedlo a bar. Pričom kategórie majú svoju hierarchiu, a tak pizza spadá pod reštaurácie rovnako ako zmrzlináreň. Jeden podnik

môže spadať pod viacero kategórií (v realite minimálne 1, medián je 4 a maximum 37). Zaujímavé je, že tzv. dlhý chvost sa uplatňuje aj v tejto analógii - väčší počet kategórií obsahuje menší počet podnikov a naopak (Obrázok 9). Niektoré kategórie majú identický názov, ale v hierarchii spadajú pod rôzne kategórie (napr. *Restaurants*, teda reštaurácie). Celkový počet kategórií je 2 464.



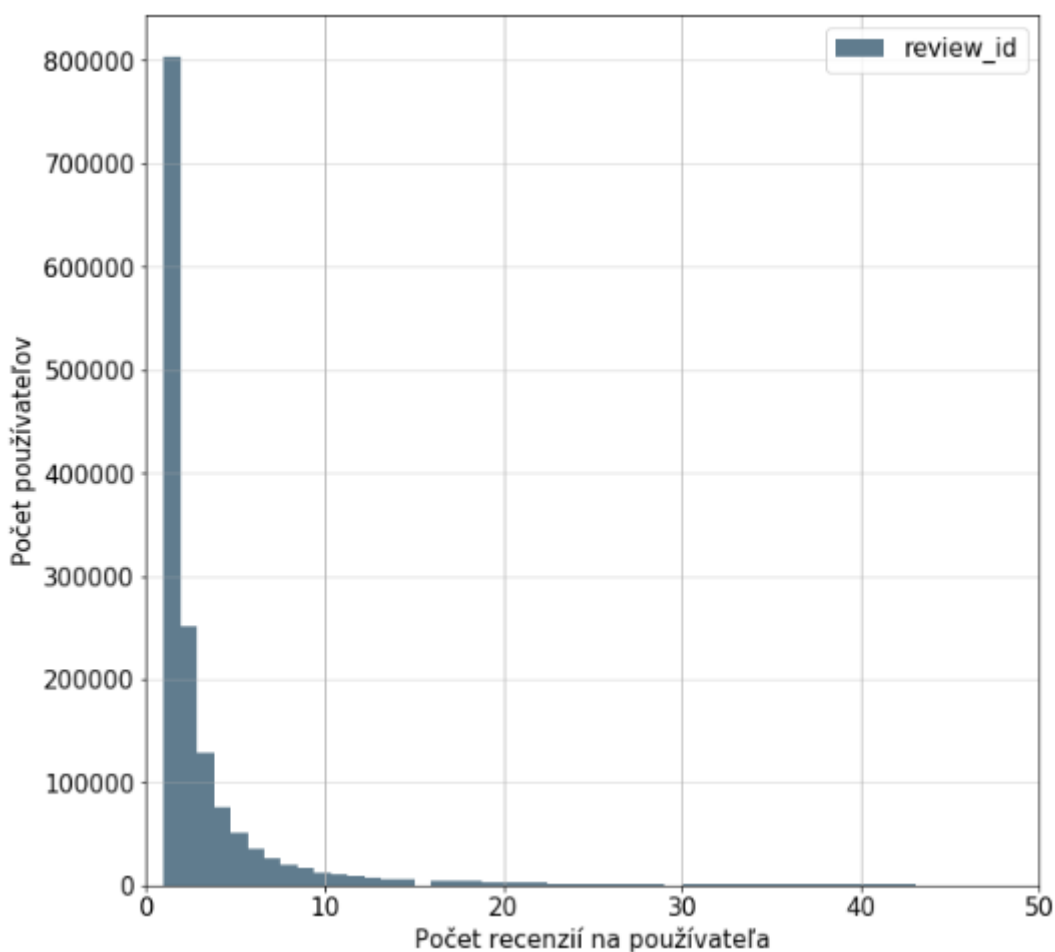
Obr. 9: Histogram kategórií podľa počtu podnikov, ktoré pod ne spadajú.

Podobne ako pri kategóriách, aj s geografickou lokalitou je to podobne – sú mestá, kde je len jeden podnik, ale maximum je takmer 29 000.

Okrem toho má každý podnik priradenú hodnotu - priemer hodnotení recenzií.

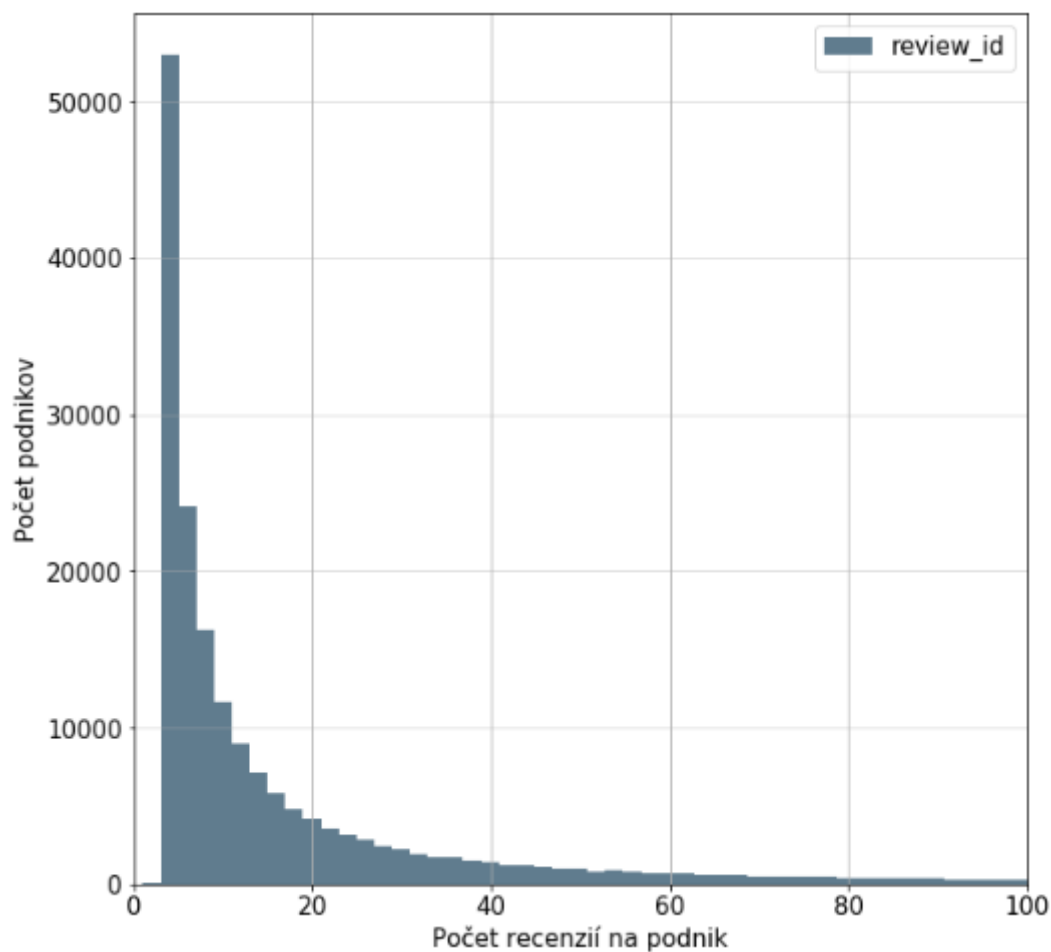
6.2.3 Recenzie

Počty recenzií na používateľa opäť pripomínajú dlhý chvost - nad 1000 recenzií napísalo 14 používateľov, menej ako 1000 ale viac ako 500 - 119 používateľov, medzi 500 a 76 - 5 913 a menej ako 76 recenzií napísalo 1 512 123 používateľov (Obrázok 10). Čiže, minimálny počet recenzií na používateľa je 1, priemer 4, maximálne viac ako 3700, ale medián je len 1.



Obr. 10: Histogram používateľov podľa počtu recenzií, ktoré nahrali.

Podobné je rozloženie počtu recenzií medzi podniky - nad 1 500 recenzií má stovka podnikov, nad 500 tisíc, medzi 500 až 300 recenzií má 1 635 podnikov, 76-300 recenzií má 13 698, ale menej ako 75 recenzií má až 172 162 podnikov (Obrázok 11). Minimálny počet recenzií pre jeden podnik je 3, maximálne 7 968, pričom medián je 9 a priemer 32.



Obr. 11: *Histogram podnikov podľa ich počtu recenzií.*

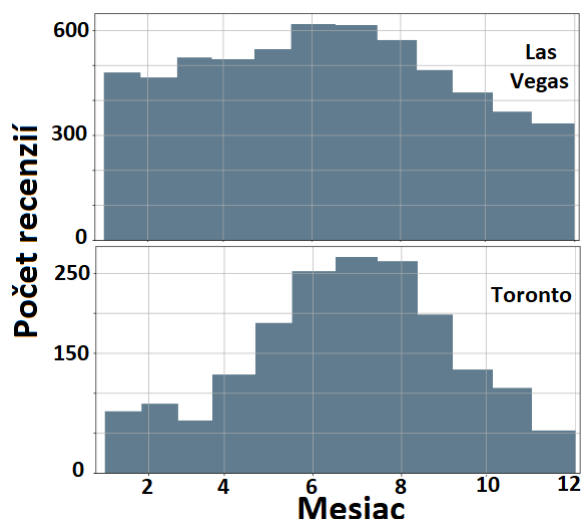
Okrem tradičných parametrov, ako je používateľ, ktorý recenziu napísal, a podnik, ktorého sa týka, obsahuje aj číselné hodnotenie (1 až 5 hviezdíček, 5 je najlepšie), dátum, textový komentár, reakcie iných používateľov - suma tých, ktorý ju označili za vtipnú, užitočnú, zaujímavú. Žiadne hodnoty atribútov nechýbajú.

6.3 Závislosť dát od časových aspektov

Na dátach sme pri bližšej analýze spozorovali niekoľko zaujímavých cyklických ale aj iných temporálnych vplyvov.

6.3.1 Cyklické temporálne vplyvy

Prvým zaujímavým zistením bolo, že niektoré kategórie podnikov sú veľmi citlivé na cyklickosť počas roka. Vplyvy ročných období, a teda vonkajšej teploty a počasia ako takého, sú pomerne známym faktom (Obrázok 12).

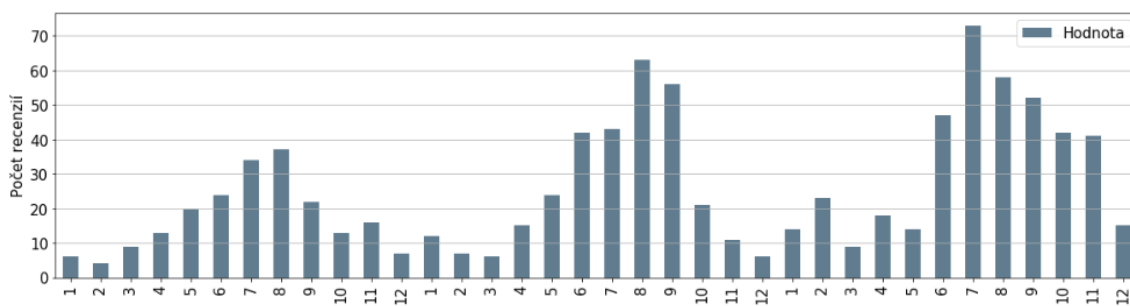


Obr. 12: Počet recenzí podľa mesiacov v priebehu roka pre podniky predávajúce zmrzlinu v mestách Las Vegas a Toronto.

Zaujímalo nás, či sa takáto cyklická sezónnosť prejaví aj na hodnotení podnikov, či menej frekventované obdobia majú lepšie alebo horšie hodnotenia, avšak ani jeden z predpokladov sa nepotvrdil - z pozorovaní nevyplýva žiadna závislosť cyklickosti a hodnotenia.

Z ďalších pozorovaní vyplynulo, že stupeň cyklickosti môže byť ovplyvnený lokalitou, o ktorú sa jedná. To znamená, že predaj zmrzliny bude ovplyvnený sezónnosťou viac v lokalite s ročnými obdobiami, resp. inak premenlivou klímou (Obrázok 12), než v oblasti relatívne stáleho počasia (Obrázok 12), teda napr. miesto, kde sa teplota stále pohybuje trvalo na úrovni letného alebo zimného počasia, ako ho vnímame v našej zemepisnej oblasti.

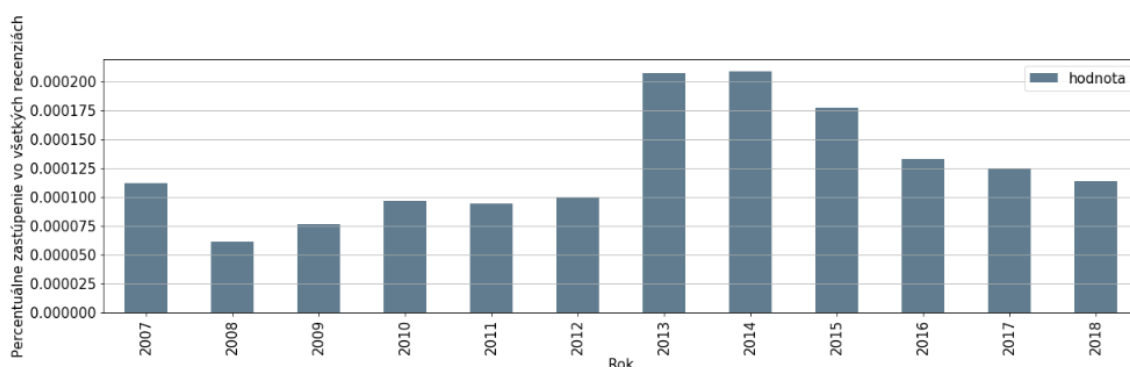
Avšak, čistý priemer počtu recenzí za jednotlivé roky podľa mesiacov môže byť zavádzajúci. Skúmali sme, či je tento vzor pozorovateľný aj pre jednotlivé roky, čo sa potvrdilo (Obrázok 13).



Obr. 13: Počet recenzí pre podniky predávajúce zmrzlinu v priebehu mesiacov niekoľkých rokov v meste Toronto.

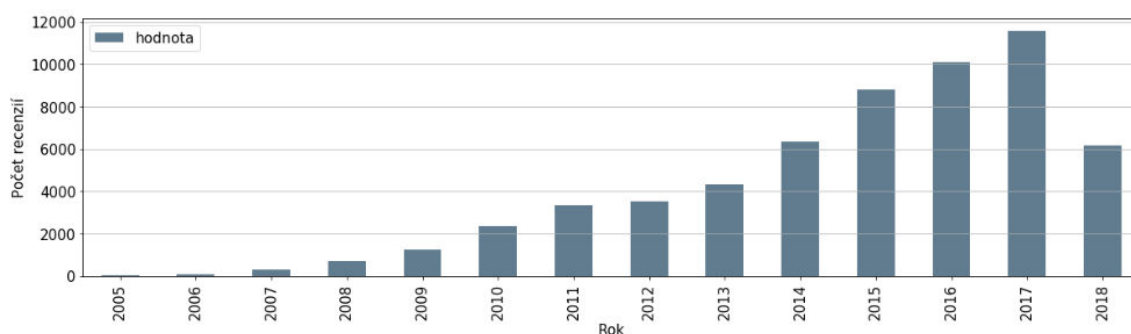
6.3.2 Necyklické temporálne vplyvy

Okrem sezónnosti, kategórie podnikov podliehali aj časovo obmedzeným trendom, teda nárastu popularity s následnými prejavmi faktoru rozpadu (Obrázok 14). Nie je to všeobecne aplikované na všetky, napr. taká kategória Pizza pokrýva s výnimkou prvého roka, kedy Yelp začínal, stabilne rovnaké percento recenzií v priebehu roka. Avšak zastúpenie iných, napr. exotických kategórií pre tradičných používateľov Yelpu (Európa, Severná Amerika, Južná Amerika, Austrália), vykazuje kontinuálnu zmenu popularity, resp. výrazne rôzne počty recenzií pre jednotlivé roky.

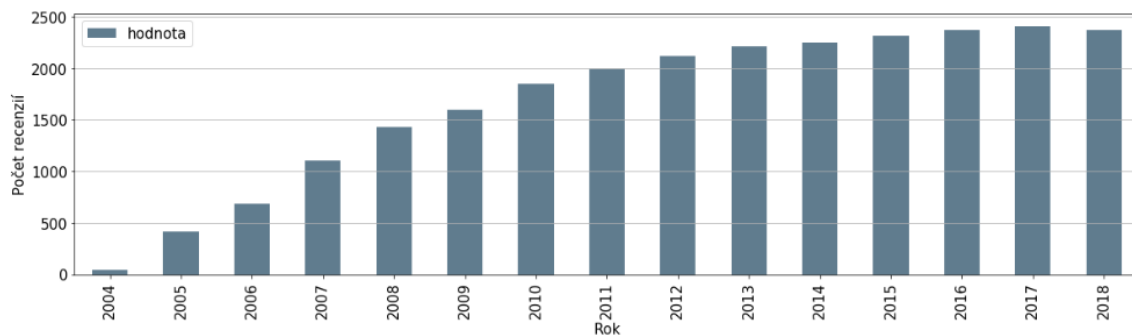


Obr. 14: Percentuálne zastúpenie recenzií pre podniky zameriavajúce sa na predaj tabaku a príslušenstva voči všetkým recenziám.

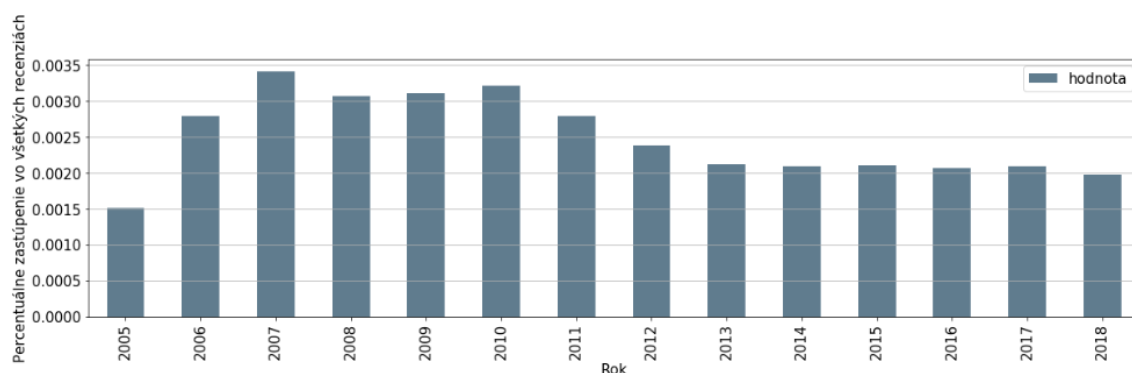
Ako príklad si môžeme zobrať suši reštaurácie, ktoré síce v počte recenzií pre jednotlivé roky (Obrázok 15) takmer kopírujú krivku celkového počtu recenzií (Obrázok 16), ale keď sa pozrieme, ako sa mení ich percentuálne zastúpenie voči všetkým recenziám (Obrázok 17), vidíme najprv postupný vzrast záujmu, následne pokles a stabilizáciu.



Obr. 15: Počet recenzií pre reštaurácie podávajúce suši v priebehu rokov.



Obr. 16: Počet všetkých pridaných recenzií podľa rokov.



Obr. 17: Percentuálne zastúpenie recenzií pre reštaurácie zameriavajúce sa na predaj suši voči všetkým recenziám.

Ako pri cyklických faktoroch, aj pri necyklických sme skúmali aj závislosť spomínaných pozorovaní s vývojom hodnotenia. Zistili sme, že zatiaľ čo keď skúmame celú kategóriu ako celok po rokoch, výrazné zmeny, až na tie, ktoré môžu byť vysvetlené počiatočnými štádiami portálu, nie sú viditeľné (Obrázok 18). Následne sme pozorovanie vykonali ešte raz, ale tentokrát pre podniky samostatne - v snahe zistiť, či zmeny hodnotenia nie sú typické skôr pre životný cyklus podniku ako takého (Obrázok 19). Konštatujeme, že tento smer preukázal väčší potenciál.

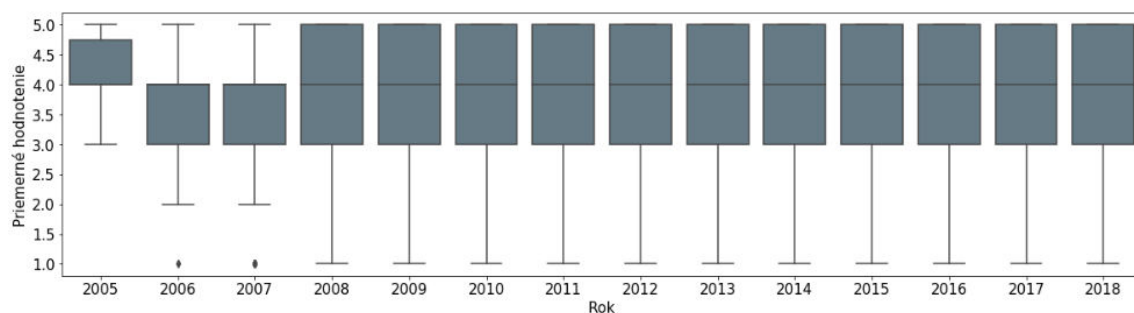
6.4 Diskusia

Na zvolenej dátovej sade z portálu Yelp sme skúmali vzťahy používateľov, podnikov a recenzií, a to primárne v oblasti temporálnych vplyvov na návštevnosť a hodnotenie podnikov. Dáta sme uvažovali v dvoch rovinách - zmeny preferencií na základe sezónnosti a zmeny popularity kategórií a podnikov a priebehu rokov.

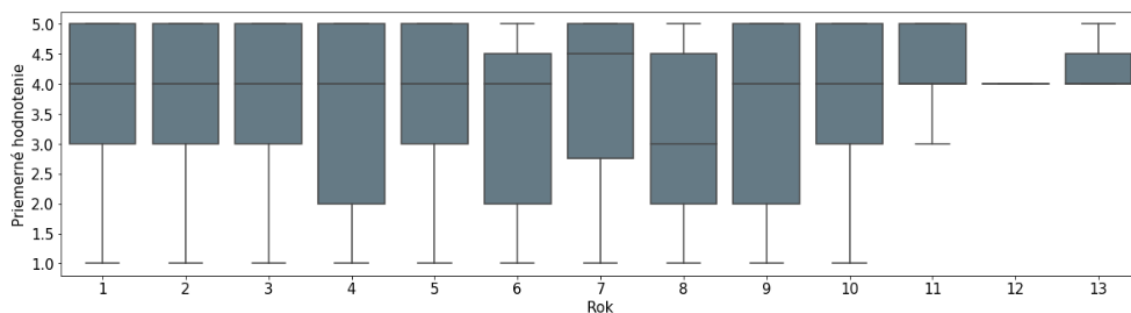
Z experimentálnych pozorovaní vyplýva, že popularita niektorých kategórií podnikov je závislá od prírodných či kultúrnych zmien počas roka, pričom dôležitá je aj intenzita

týchto zmien. Ďalšie zistenia poukazujú na výskyt krátkodobých trendov v spoločnosti ako celku, teda že niektorá kategória môže prudko vzrásť a následne padnúť v ochote návštev nových používateľov. Zaujímavým problémom na ďalšie skúmanie je aj hodnotenie podnikov na základe dĺžky ich prevádzky.

Naším ďalším cieľom je overiť, či sa tieto pozorovania odzrkadlia aj v praxi pri odporúčaní, ak sa zohľadnia príslušné črty.



Obr. 18: Hodnotenie podnikov kategórie suši podľa rokov.



Obr. 19: Hodnotenie podnikov kategórie suši podľa roku prevádzky.

7 Overenie navrhnutých metód odporúčania s temporálnymi aspektami

Okrem výberu správneho datasetu pre overenie metódy odporúčania s temporálnymi aspektami je potrebné určiť aj vhodný spôsob implementácie. Dôležité je pritom myslieť na špecifické vlastnosti odporúčania POI, a to napríklad nadpriemernú riedkosť dát.

7.1 Realizácia metódy

Pre účely odporúčania bolo v priebehu rokov vytvorených niekoľko knižníc, tu je ich stručný, neúplný prehľad, zameraný na tie použiteľné v jazyku Python:

- mREC¹⁵
 - možnosť predspracovania, rôznych metrík vyhodnotenia,
 - umožňuje rôzne modely, napríklad SLIM (lineárna regresia), WRMF (váhovaná maticová faktorizácia), hybridné odporúčanie,
 - podporuje aj paralelné odporúčanie.
- LightFM¹⁶
 - kolaboratívne metódy, prípadne kombinácia s obsahom,
 - možnosť implicitnej aj explicitnej spätnej väzby,
 - dokáže spracovať pozitívnu aj negatívnu spätnú väzbu,
 - výhodou je, že knižnica má za sebou už roky vývoja,
 - je možné použiť faktorizáciu matíc.
- Surprise¹⁷
 - aj odporúčanie, aj práca s dátami (napr. matice alebo krížová validácia),
 - kvalitná dokumentácia,
 - dobré podmienky na vytvorenie vlastného odporúčacieho systému,
 - medzi podporované funkcie patria metódy založené na susednosti (vrátane faktorizácie matíc v podobe SVD a SVD++) aj rôzne možnosti vyhodnotenia (kosínusova aj Pearsnova vzdialenosť a iné),

¹⁵<https://github.com/Mendeley/mrec>

¹⁶<https://github.com/lyst/lightfm>

¹⁷<http://surpriselib.com/>

- ak sa pre používateľa nedá odporúčať, vyhodí výnimku.
- Flurs¹⁸
 - vhodný pre prúdové dáta,
 - dokáže zohľadniť v odporúčaní aj kontext (udalosť môže mať kontext),
 - umožňuje faktorizáciu matíc,
 - knižnica je ešte len na začiatku vývoja.
- Alpenglow¹⁹
 - rámec založený na C++, ale s použiteľným Python API,
 - je vhodný najmä na online učenie,
 - dá sa použiť napríklad na nepersonalizovanú dočasnú popularitu a odporúčanie položka-položka, je obzvlášť vhodný pre metódy susednosti zohľadňujúce časové aspekty,
 - je použiteľný aj na faktorizáciu matíc.

Keďže plánujeme použiť kolaboratívne metódy v kombinácii s faktorizáciou matíc, pritom založenú na explicitnej spätnej väzbe, rozhodli sme pre hybridnú knižnicu LightFM [34]. Tá okrem iného umožňuje prepojenie kolaboratívneho prístupu s odporúčaním založeným na obsahu. Pritom modeluje používateľov a položky ako lineárnu kombináciu latentných faktorov ich črt (metadát, ktoré sú základnou reprezentáciou používateľa a položky). Latentná reprezentácia používateľa alebo položky je súčet jeho latentných vektorov, odchýlka sumou odchýlok črt. Každá črta je reprezentovaná aj skalárnou odchýlkou pre jednotlivých používateľov a položky.

Informácie sú teda reprezentované troma maticami:

- Matica interakcií - $U \times I$ - vyjadruje transakcie a hodnotenie položku používateľom (pozitívne aj negatívne).
- Matica črt používateľov - $U \times \text{črty } U$ - stĺpce reprezentujú obdoby črt, prieniky s používateľmi ich hodnotu, napr. normalizovanú na binárnu.
- Matica črt položiek - $I \times \text{črty } I$ - obdobne.

Finálne rozhodnutie o odporúčaní je z väčšej časti založené na základe skalárneho súčinu latentných reprezentácií používateľa a položky a upravené odchýlkami črt - či už

¹⁸<https://github.com/takuti/flurs>

¹⁹<https://github.com/rpalovics/Alpenglow>

používateľa, položky alebo oboch (nemusia byť dané oboje, dokonca sa dá vyhodnocovať aj úplne bez nich).

V súlade s návrhom metódy odporúčania (Kapitola č. 5) sme ako parametre LightFM zvolili stratovú funkciu WARP a program hodnoteného učenia adadelta. Počet komponentov (dimenzionalita latentných vnorení črt, 30) a počet epôch (od 265 po 600 v závislosti od veľkosti datasetu) sme experimentálne zvolili pre každú dátovú sadu zvlášť nad základným odporúčaním. Počet epôch predstavuje prechod dát modelom a teda jeho iteráciu učenia sa. Je potrebné zvoliť si počet epôch pre každú dátovú sadu samostatne. Ostatné parametre nie sú v kontexte experimentu podstatné.

Pokiaľ črty používateľov a položiek nie sú dodané, vykoná sa základné kolaboratívne filtrovanie s využitím faktorizácie matic²⁰. Pri odporúčaní sa kladie dôraz na samotný výskyt budúceho hodnotenia, nie na predikciu jeho hodnoty.

LightFM používame v kombinácii s knižnicou `scipy.sparse`²¹, ktorá slúži na prácu s nadrozmernými, ale riedkymi maticami.

Z dôvodu potreby analýza dát a vizualizácií využívame aj ďalšie špeciálne knižnice pre jazyk Python (verzia 3.6), ako napríklad `pandas`²² (dátová analýza, unikátne dátové štruktúry), `matplotlib`²³ (2D vykresľovanie), `seaborn`²⁴ (štatistická dátová vizualizácia, založené na `matplotlib`), `numpy`²⁵ (vedecké výpočty, napr. lineárna algebra) a `datetime`²⁶ (základné typy pre dátum a čas).

Ako vývojové prostredie slúži Jupyter Notebook na platforme Anaconda. Pomocou zvoleného jazyka a knižníc je možné spracovať dátovú sadu a vykonať samotné odporúčanie.

7.1.1 Spracovanie datasetu a výber podmnožiny dát

Pri práci s Yelp datasetom je ako prvé potrebné dáta konvertovať - originál má formát JSON-u a pomocou upraveného skriptu²⁷ priamo od danej spoločnosti (originál je ale potrebné upraviť na aktuálnu verziu Python, keďže skript nefunguje pre verzie Pythonu 3 a vyššie) sa dá konvertovať na CSV súbory.

Počas prvých experimentov sme sa z dôvodu výpočtovej náročnosti rozhodli zobrať len podmnožinu recenzií z jedného mesta, konkrétne North Las Vegas (Tabuľka č. 2). Ten zahŕňa 26 641 používateľov, 1 508 podnikov a 44 800 recenzií (nevyskytujú sa dup-

²⁰https://lyst.github.io/lightfm/docs/_modules/lightfm/lightfm.html#LightFM.predict

²¹<https://docs.scipy.org/doc/scipy/reference/sparse.html>

²²<https://pandas.pydata.org/>

²³<https://matplotlib.org/>

²⁴<https://seaborn.pydata.org/>

²⁵<http://www.numpy.org/>

²⁶<https://docs.python.org/2/library/datetime.html>

²⁷<https://github.com/Yelp/dataset-examples>

Tabuľka 2: *Prehľad dátových sád použitých pre tréning a validáciu riešenia.*

Označenie	Mesto	Počet POI	Počet recenzií	Počet používateľov
A	North Las Vegas	1142	11700	1269
B	Markham	1602	27408	1794
Spolu		2744	39108	3063

Tabuľka 3: *Prehľad dátových sád použitých pre overenie riešenia.*

Označenie	Mesto	Počet POI	Počet recenzií	Počet používateľov
X	Mississauga	2636	25172	1746
Y	Champaign	1132	14088	1207
Silné letá/zimy	Spolu	3768	39260	2953
Z	Goodyear	692	7669	774
W	Glendale	2718	28444	2797
Mierne zimy	Spolu	3410	36113	3571
XYZW	SPOLU	7178	75373	6524

licity, teda rovnaká dvojica používateľ-podnik). Rozhodli sme sa vybrať len používateľov, ktorý napísali aspoň päť recenzií, čo obmedzilo množinu na 1 269 používateľov, 1 142 podnikov a 11 700 recenzií. Dáta ale simulujú reálne podmienky, lebo neurčujeme, koľko majú používatelia transakcií v testovacej a trénovacej množine, teda sa vyskytujú aj noví používatelia, i skúsení. Chýbajúce hodnoty atribútov nedoplňame, lebo zatiaľ využívame len tie s dostatočným počtom výskytov.

Keďže sledujeme aj vplyv sezónnosti závislej od lokality na výsledky odporúčania, vytvorili sme obdobný dataset aj z dát z mesta Markham (Tabuľka č. 2), ktoré má severnejšiu polohu, avšak veľkosť dát (14 059 používateľov, 1 699 podnikov, 45 247 recenzií) sa podobá prvej vybranej podmnožine.

Pre následné overenie postupov sme, ako je v súlade s návrhom (Kapitola č. 5), zvolili ďalšie dátové sady (Tabuľka č. 3). Jedná sa o mestá Mississauga, Champaign, Goodyear a Glendale, každé z iného štátu.

Keďže testujeme časové aspekty a to z dvoch pohľadov, je potrebné vytvoriť pre overenie špecifické sady trénovacích a testovacích dát, na ktorých sa budú dať temporálne aspekty vyhodnocovať. Podľa návrhu teda pre každé zo zvolených oblastí vznikli trénovacia sada a testovacia sada. (V prípade sezónnosti po 12 sád - ako modelovú situáciu si môžeme predstaviť portál, ktorý vždy na začiatku mesiaca posiela email s odporúčaniami svojim zákazníkom). Testovacie sady začínajú vždy júlom 2017.

V prípade sezónnosti končia o rok neskôr. Testovacia sada pokrýva približne 20-25 % dát. Pri dlhodobých trendoch sa začínajú v júli a končia v decembri - keďže odporúčame na základe temporálnych aspektov predchádzajúcich rokov, chceli sme simulovať inú situáciu ako ideálny prípad, že odporúčame v januári, ale zároveň skončiť v rovnaký rok, aby

nemuselo byť črt dvojnásobne viac. Testovacia sada pokrýva vždy približne 10-15 % dát.

Mnohé práce [58] poukázali na potrebu postupného zapájania komponentov nových metód do odporúčacieho procesu, aby sa dalo jednoznačne určiť, ktorá zložka má aký prínos. Preto sme sa tiež rozhodli vykonať najprv základné odporúčanie bez kontextu, následne zapojiť temporálne aspekty, otestovať prínos geografického po-filtra a nakoniec kontextuálne prístupy prepojiť.

7.2 Metodológia experimentálneho overenia

Nad spracovanými dátami sme vykonali niekoľko experimentov pre overenie našich hypotéz v súlade s návrhom metód v Kapitole č. 5. V tejto kapitole sa už zaoberáme len nadstavbami nad základným odporúčaním, ktoré bolo opísané podrobne v návrhu metódy a bolo vykonané podľa plánov. Ak nastala v niektorom z experimentov výnimka, je to spomenuté pri konkrétnom príklade. Pre po-filtrovanie berieme z maticovej faktorizácie vždy prvých 20 odporúčaní, keďže to sa osvedčilo ako dostatočné množstvo pri prípadnej potrebe vyhodenia podnikov z poradia (napr. ak k nim nebolo možné vypočítať skóre, pretože žiadna ich kategória nie je dostatočnej veľkosti). Pri všetkých rozhodnutiach sa brali v úvahu oba datasets (A,B) a bolo zároveň sledované pretrénovanie, avšak nevyskytli sa žiadne kritické hodnoty.

Najprv sme zdokonaľovali metódy na základe dátových sád A (North Las Vegas) a B (Markham), následne overovali na kombináciách finálnych dátových sád X, Y, Z a W.

Prehľad používaných skratiek v tejto kapitole:

- A, B, X, Y, Z, W a ich kombinácie - vybrané a opísané dátové sady podľa Tabuliek 2 a 3 a ich kombinácie,
- MF - základná maticová faktorizácia,
- k (1, 3, 5, 10) - počet k odporúčaní, ktoré sa berie v úvahu,
- PO-F - po-filtrovanie,
- PRED-F - pred-filtrovanie,
- geo - geografické aspekty,
- temp - temporálne aspekty.

7.2.1 Zpracovanie geografických aspektov

Pre otestovanie významu geografického po-filtrovaní sme použili dataset A (North Las Vegas), avšak dáta sme výnimočne rozdelili pre každého používateľa zvlášť na tréningovú a

Tabuľka 4: *Odporúčanie s využitím maticovej faktorizácie pred a po jeho rozšírení o geografické po-filtrovanie.*

Metóda	MF		MF + geo. PO-F (d=0,1 km)	
k	Presnosť	Pokrytie	Presnosť	Pokrytie
1	73,84	10,18	99,13	13,91
3	69,56	28,33	98,63	41,48
5	65,00	43,43	92,45	63,42
10	71,81	65,31	84,71	76,66

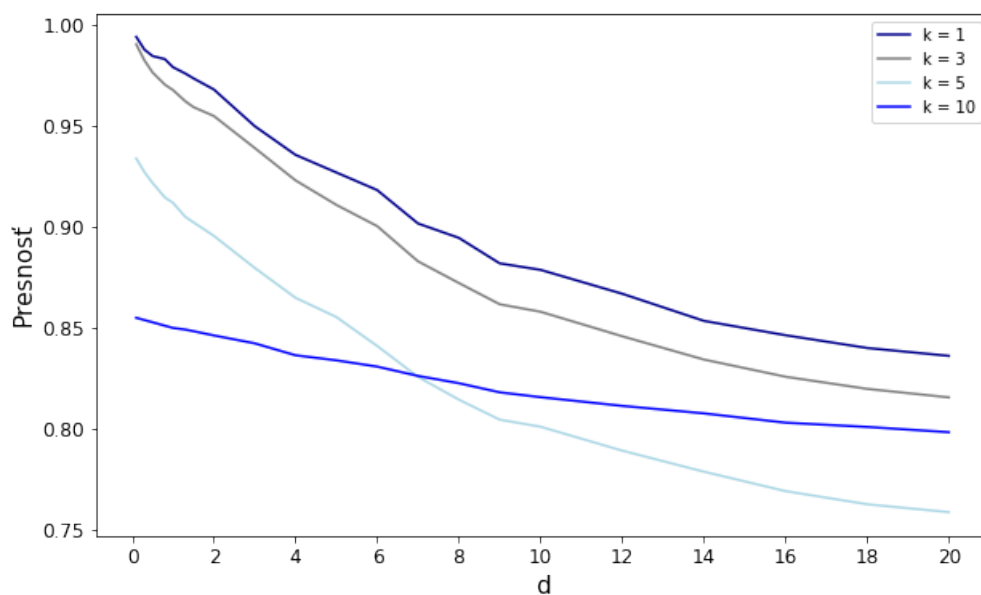
testovaciu množinu v pomere 80:20. Zmena nastala z dôvodu, že v tomto prípade nechceme simulovať temporálny vplyv a chceme čo najviac eliminovať vplyv studeného štartu, aby sme skutočne testovali len geografický vplyv.

Ako bolo opísané v Kapitole č. 5, z generovaných odporúčaní uprednostňujeme tie, ktoré sú v dosahu d od už navštívených POI. Hodnotu rádiusu d sme experimentálne určili z množiny hodnôt (0,1; 0,3; 0,5; 0,8; 1,0; 1,3; 1,5; 2,0; 3,0; 4,0; 5,0; 6,0; 7,0; 8,0; 9,0; 10,0; 12,0; 14,0; 16,0; 18,0; 20,0) km na základe dosiahnutej presnosti a pokrytia (Obrázky 20 a 21). Z týchto hodnôt sa najviac osvedčil rádius 0,1 km.

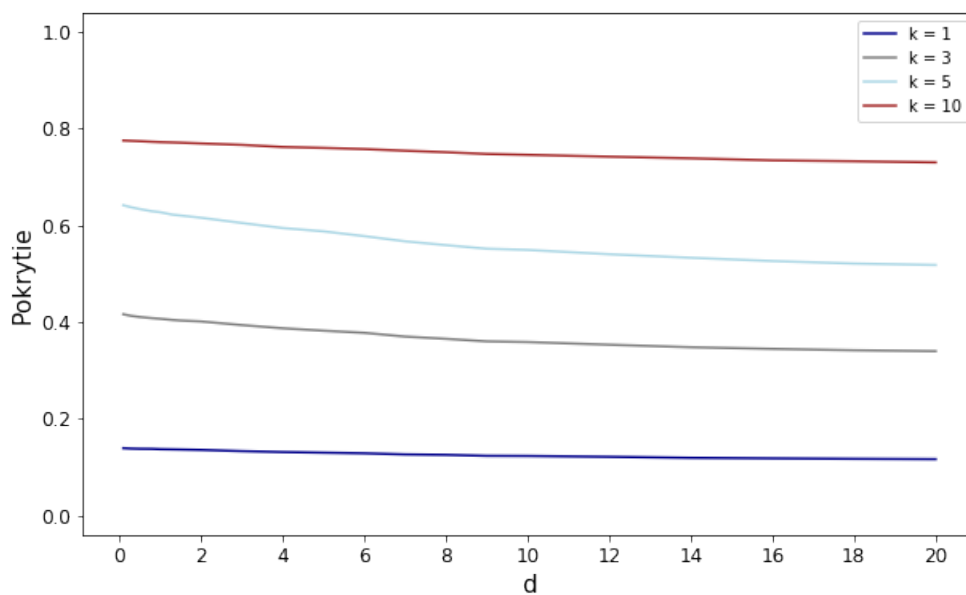
Takáto vzdialenosť, 0,1 km sa môže zdať extrémne nízka. Avšak treba si uvedomiť, že hlavný geografický filter na blízke lokality je zakomponovaný už v samotnom kolaboratívnom filtrovaní – používatelia sú podobní na základe svojich predchádzajúcich POI a na základe toho sú generované nové. Je teda nízko pravdepodobné, že by používateľovi bola vygenerovaná položka mimo vzdialenejšieho dosahu ako sú ochotní vycestovať jemu podobní používatelia. Preto filtre v rozsahu niekoľkých kilometrov sú zanedbateľné, už ošetrené kolaboratívnym prístupom.

Do hry ale vstupujú často navštevované časti mesta používateľom, napríklad okolie domova, pracoviska, obľúbeného nákupného strediska. Je pravdepodobné, že používateľ pri výbere nového POI zohľadní jeho blízkosť ku vlastným frekventovaným oblastiam. Preto v minulosti (už navštívené POI, trénovacia množina) aj budúcnosti (POI v testovacej množine) navštevuje POI vo vzájomnej bezprostrednej blízkosti.

Konkrétne zlepšenie metrík na spomínanej dátovej sade s geografickým po-filtrom s hodnotou rádiusu 0,1 je možné vidieť v Tabuľke 4. Neskoršie výsledky experimentov zahŕňajúcich aplikáciu temporálneho vplyvu vykazujú po kombinácii s geografickým po-filtrom podobné zlepšenie metrík.



Obr. 20: Závislosť presnosti@k na rádiuse d pri geografickom po-filtrovaní.



Obr. 21: Závislosť pokrytia@k na rádiuse d pri geografickom po-filtrovaní.

7.2.2 Temporálne modelovanie lokálnej sezónnosti s temporálnym pred-filtrom a geografickým po-filtrom

Vykonalí sme základné odporúčanie na dátovej sade A a B, následne sme doň pridali maticu identity a potom jednotlivo testovali všetky navrhnuté črty samostatne (dokopy

Tabuľka 5: Porovnanie základnej maticovej faktorizácie pred a po pridaní pred-filtra na základe mesiacov a následnom pridaní ďalšej črty - vážený priemer pre kategórie POI za najbližšie tri mesiace.

Dáta	Markham					
Metóda	MF		Pred-F		VP3 + pred-F + mesiac	
k	Presnosť	Pokrytie	Presnosť	Pokrytie	Presnosť	Pokrytie
1	1,04	0,51	1,4	1,21	1,47	1,26
3	1,80	1,61	2,7	2,7	2,83	2,83
5	2,55	2,50	3,33	3,33	3,5	3,5
10	4,04	4,03	3,72	3,72	4,14	4,14

6 črt). Pritom tieto črty sú rátané len pre kategórie, ktoré majú nadpriemerné výkyvy čo sa mesiacov týka (ako pozitívna hodnota) alebo výrazne nižšie (negatívna hodnota). Ostatné sú pri generovaní hodnôt ignorované. Prospešnosť tohto prístupu sa prejavila pri experimentovaní s po-filtrovaním na základe sezónnosti.

Zvlášť sme otestovali aj temporálny pred-filter. Najlepšie pre oba datasets sa osvedčil práve pred-filter a pri modelovaní črta reprezentujúca vážený priemer kategórií POI pre 3 mesiace (súčasný a dva susedné). Výsledky pred-filtra sa ešte zlepšili pridaním črty mesiaca (testovali sa obe formy, cyklicky kódovaný ako sínus a kosínus sa ukázal mierne vhodnejší). Vytvorila sa teda konečná kombinácia pred-filtra, modelovania črt (mesiac cyklicky, vážený priemer pre tri mesiace). Prehľad výsledkov je v Tabuľke 5.

Aj keď samotné *one hot* kódovanie kategórií malo podobné výsledky, uprednostňujeme numerické črty z dôvodu redukcie počtu atribútov a teda vyvarovaniu sa pred pretrénovaním.

Podrobnejšie výsledky pre jednotlivé črty sú v Prílohe D.

Tabuľka č. 6 zobrazuje porovnanie úspešnosti odporúčania na dátových sadách A a B pred a po zohľadnení kontextu. Je očividné, že lepšie výsledky dosahuje v Markhame, než v North Las Vegas, kde ani sezónne výkyvy počasie nie sú veľmi znateľné.

7.2.3 Temporálne modelovanie lokálnych trendov s geografickým po-filtrom

Aj pri lokálnych trendoch sa postupne vyskúšali všetky navrhnuté sady črt - najlepšie obstála trojica vyjadrujúca aritmetický priemer pre kategórie - rozdiel počtu recenzií oproti predchádzajúcemu roku (tri črty pre posledné tri roky). Keď sme sa pokúšali použiť kombináciu tejto trojice do jedného skóre, ktoré sa používa pri po-filtrovaní, úspešnosť bola podstatne nižšia. Okrem tejto črty sme samozrejme použili aj črtu štátu (Tabuľka č. 7). Podrobnejšie výsledky pre jednotlivé črty sú v Prílohe D.

Tabuľka č. 8 zobrazuje porovnanie úspešnosti odporúčania na dátových sadách A a B pred a po zohľadnení kontextu. V oboch prípadoch dosahuje metóda zlepšenia.

Tabuľka 6: *Odporúčanie s využitím maticovej faktorizácie pred a po jeho rozšírení o temporálne prefiltrovanie, modelovanie na základe črty mesiac (cyklicky vyjadrený), črty štát a črty vážený priemer pre kategórie za tento, predchádzajúci a nasledovný mesiac, spolu s geografickým po-filtrovaním.*

Dátová sada	North Las Vegas			
Metóda	MF		Temp. modelovanie + pred-F + geo. po-F	
k	Presnosť	Pokrytie	Presnosť	Pokrytie
1	0,4	0,27	0,44	0,44
3	0,63	0,56	0,62	0,44
5	0,82	0,79	0,79	0,79
10	1,29	1,29	0,79	0,79
Dátová sada	Markham			
1	1,04	0,51	1,38	1,26
3	1,80	1,61	3,29	3,29
5	2,55	2,50	3,84	3,84
10	4,04	4,03	4,33	4,33

Tabuľka 7: *Porovnanie základnej maticovej faktorizácie pred a po obohatení o črty položiek - štát a trojica vyjadrujúca aritmetický priemer pre kategórie - rozdiel počtu recenzií oproti predchádzajúcemu roku (tri črty pre posledné tri roky).*

Dátová sada	Markham			
Metóda	MF		Temp.črta, štát	
k	Presnosť	Pokrytie	Presnosť	Pokrytie
1	1,93	0,45	5,32	1,12
3	2,95	1,89	4,07	2,45
5	3,47	2,90	4,49	3,58
10	5,41	5,19	7,47	7,08

Tabuľka 8: *Odporúčanie s využitím maticovej faktorizácie pred a po jeho rozšírení o temporálne modelovanie na základe črty pre štát a trojice vyjadrujúcej aritmetický priemer pre kategórie - rozdiel počtu recenzií oproti predchádzajúcemu roku (tri črty pre posledné tri roky). Na konci je vykonané geografické po-filtrovanie.*

Dátová sada	North Las Vegas			
Metóda	MF		Modelovanie a geo po-F.	
k	Presnosť	Pokrytie	Presnosť	Pokrytie
1	1,78	0,43	2,22	0,80
3	1,56	1,09	2,22	1,95
5	2,52	2,3	2,57	2,44
10	3,14	3,13	3,74	3,73
Dátová sada	Markham			
1	1,93	0,45	5,19	1,08
3	2,95	1,89	4,16	2,54
5	3,47	2,90	4,49	3,56
10	5,41	5,19	7,63	7,24

7.2.4 Temporálny po-filter lokálnej sezónnosti s temporálnym pred-filtrom a geografickým po-filtrom

Použili sa všetky prístupy spomínané v návrhu, pričom tentokrát už nemusíme opakovať samostatný temporálny pred-filter, keďže výsledky by boli rovnaké, ako pri modelovaní. Ako najšťubnejšie sa ukázali vyvážené spriemerované kategórie s váženou kombináciou s pôvodným poradím (len aktuálny mesiac, pridanie okolitých nepomáhalo), následne tie s rovnakým skóre a neváženou kombináciou poradia (Tabuľka č. 9). Pomery poradí sa taktiež stali objektom optimalizácie.

Pri porovnaní s po-filtrovaním na základe popularity je však vidno, že aj táto najlepšia metóda ledva obstojí. Avšak, môžeme zhodnotiť, že výsledky predchádzajúceho prístupu, modelovania, porážajú aj preusporiadanie na základe popularity.

Experimentovaním sme však spravili dva zaujímavé objavy. Prvým je, že tieto črty je prospešnejšie rátať len pre kategórie, ktoré majú nadpriemerné výkyvy, čo sa mesiacov týka (ako pozitívna hodnota) alebo výrazne nižšie (negatívna hodnota). Ostatné sú pri generovaní hodnôt ignorované.

Skúmali sme aj, aký má sezónnosť vplyv na kvalitu odporúčania vzhľadom na predchádzajúci počet transakcií daného človeka. Zistili sme, že sezónnosť nie je vhodná pre riešenie studeného štartu, ako by sa na prvý pohľad mohlo zdať. Pri tom ju vždy predbehol po-filter na základe popularity. Čím ale používatelia boli viac expertní, popularita strácala vplyv a sezónnosť porážala ako základné odporúčanie tak i popularitu. Pravdepodobne lebo pre novických a príležitostných používateľov sú lákadlom len globálne trendy, zatiaľ čo na dlhodobých používateľoch, ktorí nevyužívajú len najpopulárnejšie položky, je

Tabuľka 9: Porovnanie základnej maticovej faktorizácie pred a po pridaní po-filtra na základe sezónnosti - vážený priemer pre kategórie POI (VK) a následne vyvážená (VP) alebo nevyvážená (NP) kombinácia poradia s pôvodným.

Dátová sada	Markham			
	MF		Popularita po-F.	
Metóda	Presnosť	Pokrytie	Presnosť	Pokrytie
k				
1	1,08	0,51	0,98	0,44
3	1,27	1,09	1,56	1,34
5	2,73	2,67	2,15	2,08
10	4,28	4,28	4,53	4,52
Metóda	VK, VP		VK, NP	
1	1,24	0,46	0,74	0,29
3	1,48	1,33	1,46	1,30
5	2,37	2,34	2,70	2,67
10	4,70	4,70	4,60	4,60

vidieť skutočné vplyvy sezónnosti. Nárast vplyvu sezónnosti bol úmerný počtu transakcií používateľa, zatiaľ čo vplyv popularity je nepriamo úmerný.

Z Tabuľky č. 11 je zrejmé, že vplyv sezónnosti je tak ako pri modelovaní, aj pri po-filtrovaní v North Las Vegas výrazne nižší, až negatívny, oproti Markhamu.

7.2.5 Temporálny po-filter lokálnych trendov s geografickým po-filtrom

Na základe každej sady (trojice) črt si vyrátame podľa uvedeného vzťahu skóre, podľa ktorého preusporiadame odporúčania. Dokopy nám tak vzniklo šesť metód, k tomu ich porovnávame ešte s čistou popularitou. Pritom opäť optimalizujeme vyváženie kombinácie pôvodného a nového poradia. Hoci jedna z metód má dobré výsledky oproti základnej faktorizácii, nestačí na po-filtrovanie na základe popularity (Tabuľka č. 11). Avšak modelovanie porazilo po-filtrovanie na základe popularity aj v prípade tejto hypotézy, teda pri dlhodobých trendoch.

Uspokojivé výsledky sa nám podarilo dosiahnuť len pre dátovú sadu B, Markham (Tabuľka č. 12).

7.2.6 Finálny výber metódy a jej overenie

Po-filtrovanie síce prinášalo pridanú hodnotu k základnej maticovej faktorizácii, ale málo kedy vedelo poraziť jednoduchý po-filter na základe celkovej popularity. Naopak, modelovanie dosahuje voči oboch hypotézam uspokojivé výsledky a pri porovnaní s po-filtrovaním na základe popularity môžeme zhodnotiť, že porazilo aj túto metódu. Okrem toho prináša výhodu rýchlosti, ktorá je oproti po-filtrovaniu niekoľkonásobná.

Tabuľka 10: Porovnanie základnej maticovej faktorizácie pred a po pridaní po-filtra a pred-filtra na základe sezónnosti a geografického po-filtra. Porovnanie na dvoch dátových sadách - North Las Vegas a Markham.

Dátová sada	North Las Vegas			
Metóda	MF		Temp. po-F, pred-F, geo po-F	
k	Presnosť	Pokrytie	Presnosť	Pokrytie
1	0,00	0,00	0,44	0,44
3	1,39	1,39	0,44	0,44
5	1,39	1,39	0,44	0,44
10	2,54	2,54	1,18	1,18
Dátová sada	Markham			
1	1,08	0,51	1,27	1,14
3	1,27	1,09	3,24	3,24
5	2,73	2,67	3,84	3,84
10	4,28	4,28	4,34	4,34

Tabuľka 11: Porovnanie základnej maticovej faktorizácie pred a po pridaní po-filtra na základe temporálnych trendov (normalizovaný nárast/pokles oproti minulej sezóne, aritmetický priemer pre kategórie) alebo popularity. Porovnanie na dátovej sade B Markham.

Dátová sada	Markham					
Metóda	MF		Temp.črta, štát		Popularita	
k	Presnosť	Pokrytie	Presnosť	Pokrytie	Presnosť	Pokrytie
1	1,93	0,45	2,70	0,67	3,11	0,94
3	2,95	1,89	3,22	2,08	3,49	2,23
5	3,47	2,90	3,71	3,13	4,63	3,93
10	5,41	5,19	5,91	5,71	6,45	6,21

Tabuľka 12: Porovnanie základnej maticovej faktorizácie pred a po pridaní temporálneho po-filtra, na základe temporálnych trendov, a geografického po-filtra. Porovnanie na dvoch dátových sadách - North Las Vegas a Markham.

Dátová sada	North Las Vegas			
Metóda	MF		Temp.po-F, geo.po-F	
k	Presnosť	Pokrytie	Presnosť	Pokrytie
1	1,78	0,43	0,00	0,00
3	1,56	1,09	0,67	0,49
5	2,52	2,30	0,77	0,72
10	3,14	3,13	2,46	2,46
Dátová sada	Markham			
1	2,70	0,64	3,32	1,06
3	3,16	1,99	3,80	2,44
5	3,43	2,86	4,19	3,56
10	5,65	5,46	6,32	6,10

Tabuľka 13: Porovnanie základnej maticovej faktorizácie, lokálneho a globálneho temporálneho vplyvu sezónnosti na relevantnosť odporúčaní. Lokálny aj globálny prístup zahŕňa okrem modelovania temporálny pred-filter a geografický po-filter.

Metóda	Základná MF		Lokálne temp. črty		Globálne temp. črty	
k	Presnosť	Pokrytie	Presnosť	Pokrytie	Presnosť	Pokrytie
XYZW						
1	0,92	0,45	1,56	1,17	1,22	0,94
3	1,43	1,31	1,33	1,33	1,21	1,21
5	1,64	1,61	1,33	1,33	1,21	1,21
10	2,01	2	1,33	1,33	1,21	1,21
XY						
1	1,97	0,46	2,89	2,11	2,65	2,09
3	1,21	0,99	2,54	2,54	2,32	2,32
5	1,45	1,35	2,63	2,63	2,41	2,41
10	1,92	1,91	2,82	2,82	2,41	2,41
ZW						
1	0,71	0,15	0	0	0	0
3	1,16	1,08	0	0	0	0
5	1,43	1,41	0	0	0	0
10	2,7	2,7	0	0	0	0

Finálne overenie hypotéz prebieha podľa návrhu metódy (Kapitola č. 5) na troch dátových sadách, pričom na každej sa vykonajú tri výpočty - základná maticová faktorizácia, jej obohatenie o lokálne temporálne vplyvy a o globálne temporálne vplyvy. Za črty modelovania sú zvolené najúspešnejšie črty z predchádzajúcich experimentov.

H1: Lokálna sezónnosť

Ako je zrejmé z Tabuľky č. 13, pri dátovej sade, ktorá sa skladá zo štátov s výkyvmi klímy v priebehu roka v podobe ročných období je výhodné zakomponovať aj vplyvy sezónnosti (XY). Avšak, naopak, pri oblastiach bez tejto špecifikácie (ZW) to môže odporúčací systém len zbytočne zmiasť a výsledky môže výrazne zhoršiť, či už sú v podobe globálnych alebo lokálnych črt. Pravdepodobne preto sú výsledky na najpestrejšej dátovej sade XYZW síce lepšie na lokálnych temporálnych črtách, ale toto odporúčanie nie je jednoznačne lepšie oproti jeho základnej forme (je úspešnejšie pre nižšie počty vyhodnotených odporúčaní, teda pre prvé, čiastočne aj pre tri odporúčania).

Dôvody zhoršenia na sadách s menej premenlivými vonkajšími podmienkami nie sú určite spôsobené chybou implementácie alebo anomáliou náhodných faktorov (vyskytujú sa v danej knižnici), keďže sa nezmenili ani pri opakovaných pokusoch a boli vyrátané rovnakým prístupom ako nad ostatnými štyrmi dátovými sadami. Taktiež pri druhej metóde

sú výsledky dátovej sady štandardné. Extrémne zlé výsledky sú pravdepodobne spôsobené faktom, že medzi jednotlivými intervalmi (mesiacmi) nie sú dostatočné kontextuálne rozdiely, a tak vytvárajú len skreslenie a mohli by prerásť až do pretrénovania.

Môžeme skonštatovať, že uvažovanie nad regiónmi vrámci sezónnosti má význam - či už v podobe vylepšenia výsledkov v oblastiach, kde sa sezónnosť prejavuje, alebo aby sme sa jej vyhli v regiónoch, kde by mohla zbytočne uškodiť. Hypotéza je platná, lokálna sezónnosť je významnejšia ako globálna sezónnosť – v oblastiach so silnou sezónnosťou sa jej význam využije naplno, v oblastiach so zanedbateľnou sezónnosťou nebude zbytočne skresľovať výsledky.

V oblasti so sezónnosťou si môžeme efekt predstaviť na príklade používateľa, pre ktorého odporúčané POI by boli čajovňa a kaviareň s letnou terasou – kým počas letných mesiacov by bola uprednostnená kaviareň, počas chladnejších čajovňa. Avšak, pri oblastiach, kde sa celoročné podmienky nemenia, by bol podobný prístup zbytočný.

Využili sme črty reprezentujúce cyklickosť období (mesiac ako sínus a kosínus), oblasť (štát) a pravdepodobnosť pre navštívenie daného POI počas aktuálneho obdobia. Posledné bolo vyjadrené váženým priemerom hodnôt pre kategórie položky. Jednotlivé hodnoty sa vytvárajú ako pomer priemerného počtu návštev v danom období (aktuálny mesiac a jeho susedné mesiace) oproti priemernému počtu návštev za dĺžku zvoleného cyklu (rok). Tento prístup pri výbere črt zaručuje, že je metóda prenosná aj na iné dáta, prípadne iné domény.

Úspešnosť by sa pravdepodobne dala ešte zvýšiť, keby sa použitie sezónnosti uplatňovalo na skúsenejších používateľov, zatiaľ čo na tých s malým počtom transakcií by sa uplatňovalo využitie odporúčania s ohľadom na celkovú popularitu medzi položkami v lokalite používateľa.

H2: Lokálne trendy

Hypotézu 2 o vplyve lokálnych trendov sa nepodarilo dokázať (Tabuľka č. 14). Črty založené na lokálnych temporálnych vplyvoch síce dosahujú v najviac diverznej skupine lepšie výsledky ako globálne trendy, avšak nedokážu poraziť základné odporúčanie založené na maticovej faktorizácii. Ukazuje sa síce, že ak nejaké črty máme zvažovať, radšej tie v danej oblasti, avšak ich povaha musí byť odlišná od prístupu, ktorý sme zvolili.

Lepšie výsledky pritom lokálne trendy oproti globálnym dosahujú na nižšom počte odporúčaní, čo môže naznačovať, pri možnosti odporučiť extrémne nízky počet položiek používateľovi je efektívnejšie uprednostniť populárne položky v jeho oblasti, než celkovo populárne. Nie v zmysle konkrétnych POI, samozrejme, že tie uprednostní vo svojom okolí, ale v zmysle kategórií. V jednej oblasti sa môžu stávať populárnejšími kaviarne spojené s kníhkupectvami, zatiaľ čo v inej lokálne pivovary.

Tabuľka 14: Porovnanie základnej maticovej faktorizácie, lokálneho a globálneho temporálneho vplyvu trendov na relevantnosť odporúčaní. Lokálny aj globálny prístup zahŕňa okrem modelovania geografický po-filter.

Metóda	Základná MF		Lokálne temp. črty		Globálne temp. črty	
k	Presnosť	Pokrytie	Presnosť	Pokrytie	Presnosť	Pokrytie
XYZW						
1	1,64	0,65	0,85	0,32	0,24	0,08
3	1,77	1,69	0,64	0,49	0,44	0,28
5	1,92	1,88	0,98	0,86	0,45	0,39
10	2,57	2,57	2,13	2,08	1,56	1,52
XY						
1	1,46	0,18	2,63	0,81	1,17	0,26
3	2,05	1,13	1,9	1,17	1,56	0,83
5	2,29	1,75	1,59	1,32	2,16	1,82
10	3,3	3,13	2,35	2,35	3,01	2,88
ZW						
1	0,83	0,33	0,62	0,35	1,25	0,47
3	1,39	1,19	0,97	0,84	1,21	0,99
5	2,16	2,09	1,75	1,7	1,39	1,28
10	3,07	3,06	2,09	2,09	2,41	2,41

Zaujímavé je, že čím viac počtom i diverzitou oblastí sa v dátovej sade nachádza, tým sú výsledky pre lokálne trendy lepšie, ako pre globálne. Síce vybrané oblasti reprezentujú štyri rôzne štáty, avšak všetky sa nachádzajú v Severnej Amerike. Bolo by zaujímavé použiť pestrejšiu zmes štátov, nebolo to však možné na tejto dátovej sade pri objeme dát, ktorý sme chceli využiť pre každú oblasť.

Okrem toho by mohlo byť prínosné zaoberať sa ohraničením oblasti. Najprv si vytvoriť model na určenie podobností medzi blízkymi drobnými oblasťami a následne ich spojiť do väčších na jeho základe. Tieto konečné oblasti by sa použili pri odporúčaní.

7.3 Diskusia

V tejto kapitole sme opísali výber vhodných dátových sád a experimenty za účelom overenia dvoch hypotéz. Všetky experimenty boli primárne postavené na využití temporálnych črt v kontexte geografických oblastí, teda vplyv lokálnej sezónnosti a lokálnych temporálnych trendov. Práve zohľadnenie lokálnych temporálnych vplyvov robí metódu jedinečnou, keďže väčšina predchádzajúcich prác sa zaoberala len globálnymi vplyvmi. Ako metriky sme použili presnosť@k a pokrytie@k.

Podarilo sa nám ukázať, že v niektorých oblastiach skutočne má význam brať v úvahu špecifickú sezónnosť. Zvyčajne sa jedná o miesta s menej stabilnou klímou počas roka, teda tie so zmenou ročných období. Naopak, pri takto stabilných oblastiach podobný prístup kvalitu odporúčaní zhoršuje. Môžeme teda vyhlásiť, že hypotéza je pravdivá, hoci jej znenie by bolo vhodné rozšíriť na - áno, je vhodné zohľadňovať sezónnosť na základe lokality - ale to aj znamená, že v niektorých oblastiach by tieto vplyvy mali byť ignorované, zatiaľ čo v iných je potreba venovať im viac pozornosti.

Ako vhodná črta sa ukázala kategória v podobe vyjadrenia jej popularity pre jednotlivé zvažované intervaly. Je to podobný záver, ako záver Zhang et al. [62], ktorí používali odhad rozloženia popularity kategórií a výpočet skóre pre relevantnosť kategórie na dátovej sade z rovnakého portálu, i keď staršej verzii. Povaha ich úlohy bola mierne iná, takisto neuvádzajú, že by rozdeľovali dáta na 12 testovacích (čo znižuje úspešnosť odporúčania), preto je ťažké výsledky priamo porovnať, avšak nimi uvedená presnosť sa pohybuje na mierne nižších hodnotách ako naša pri finálnom vyhodnotení.

Podstatné je, že naša metóda je koncipovaná tak, aby nebola viazaná na konkrétnu dátovú sadu a dokonca by sa dala preniesť aj na iné domény.

Avšak, snáď najzaujímavejším poznatkom v tejto oblasti je zistenie, že ak sezónnosť význam má, nie pre používateľov-novícov s krátkou históriou, ale práve pre expertných používateľov, ktorí vykonali vyšší počet transakcií.

Druhú hypotézu o lokálnych trendoch môžeme prehlásiť za nepotvrdenú. Globálne a lokálne trendy mali približne rovnakú úspešnosť pre všetky dátové sady. Úspešnosť je ale vždy nižšia, než pôvodné odporúčanie.

Pri oboch hypotézach sme zvolili dva rôzne prístupy – po-filtrovanie a modelovanie. Po-filtrovanie sa ukázalo ako ťažkopádnejšie a menej efektívne, popularitu zvyčajne porazilo len pre aktívnejších používateľov. Avšak, práve to je jej prínos pre prácu, keďže vďaka jednoduchému sledovaniu temporálneho vplyvu a popularity na prezoradenie odporúčaní vieme zhodnotiť, že pri novších a menej aktívnych používateľoch, o ktorých ešte nemáme dostatočné informácie, je lepšie voliť všeobecne populárne položky.

Napriek tomu je maticová faktorizácia s kombináciou črt vhodným prístupom, ako pokračovať v skúmaní. Narozdiel od po-filtrovania je síce ťažšie vidieť do jej vnútorného

fungovania, avšak dokáže s väčšou ľahkosťou pracovať s latentnými črtami položiek a používateľov. Pri jej kombinácii s pred-filtrovaním na základe sezónnosti by bolo zaujímavé zmeniť intervaly jednotlivých sezón – keďže sa táto metóda ukázala prospešnou pri oblastiach s výraznými ročnými obdobiami, mohli by sa intervaly zvoliť podľa nich a nie podľa mesiacov.

Iným zaujímavým rozšírením by bolo zmeniť odporúčanie POI na odporúčanie kategórií POI. Súčasne sústredenie sa na skúmanie hranice, kedy sa kategória vníma ako výnimočne populárna či nepopulárna počas roka, ako stúpanie či klesanie jej popularity v priebehu rokov súvisí so samostatným vývojom platformy.

Metóda je ohrozená prípadmi, keď boli recenzie používateľom pridané až dodatočne, s odstupom času. To môže nie len skresliť tréningové dáta, ale aj znižuje presnosť vyhodnotenia, najmä pri zohľadnení sezónnosti. Bolo by zaujímavé použiť ju aj na dátovú sadu, kde sú transakciami prihlásenia.

8 Zhodnotenie

Začiatok tejto práce je venovaný bližšiemu pohľadu na odporúčacie systémy, či už z pohľadu ich významu alebo princípov fungovania. Následne sme sa bližšie pozreli na špecifické kontextové odporúčanie, najmä v oblasti časových aspektov.

Najviac priestoru sme ale venovali doméne na lokalite založených sociálnych sietí, ktoré sa špecializujú na odporúčanie bodov záujmu. Táto doména je plná nezodpovedaných otázok v oblasti kontextových vplyvov, hoci sa tejto problematike venovalo už nemálo prác. Nás zaujala téma lokálnej sezónnosti kategórií, nemenej podstatná bola aj analýza dlhodobého vývoju lokálneho trendu popularity kategórií POI.

Za pomoci získaných poznatkov sme navrhli vlastné metódy, ktorých úlohou je zlepšiť existujúce prístupy odporúčania v LBSN obohatením o zložky ako geografické, tak aj temporálne. Využili sme kontextuálne modelovanie, pre-filtrovanie aj po-filtrovanie.

Najlepšie výsledky sme dosiahli s využitím lokálnej sezónnosti zohľadnenej pomocou kontextuálneho modelovania - teda pomocou maticovej faktorizácie obsahujúcej temporálne založené obsahové črty a geografické črty položiek. Tento prístup sa ukázal obzvlášť významný pre oblasti, kde sa v priebehu roka zreteľne mení počasie na základe ročných období. Metóda má veľký potenciál využitia aj v iných doménach.

Limitáciami tohto prístupu je problematika určenia vhodných časových intervalov pre sezónu a že v prípade nových používateľov sú iné vplyvy podstatnejšie, ako sezónnosť. Preto by bolo vhodné zamerať sa v budúcnosti na jej kombináciu s inými metódami, ktoré vedia naopak zmierniť studený štart.

Dôkazom toho, že kontextuálne vplyvy budú podstatné počas nasledujúcich rokov v oblasti sociálnych sietí založených na lokalite je aj tohtoročná výzva AMC RecSys Challenge 2019²⁸, ktorá je zameraná na kontextuálne odporúčanie v doméne cestovania.

²⁸<http://www.recsyschallenge.com/2019/>

Literatúra

- [1] Gediminas Adomavicius, Ramesh Sankaranarayanan, Shahana Sen, and Alexander Tuzhilin. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Trans. Inf. Syst.*, 23(1):103–145, January 2005.
- [2] Gediminas Adomavicius and Alexander Tuzhilin. Context-aware Recommender Systems. *Proceedings of the 2008 ACM Conference on Recommender Systems*, pages 335–336, 2008.
- [3] Charu C. Aggarwal. *Recommender Systems - The Textbook*, volume 40. Springer International Publishing, Cham, 2016.
- [4] Vito W Anelli, Vito Bellini, Tommaso Di Noia, Wanda La Bruna, Paolo Tomeo, and Eugenio Di Sciascio. An analysis on time- & session-Aware diversification in recommender systems. *[UMAP2017]Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 270–274, 2017.
- [5] Jie Bao, Yu Zheng, and Mohamed F. Mokbel. Location-based and preference-aware recommendation using sparse geo-social networking data. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems, SIGSPATIAL '12*, pages 199–208, New York, NY, USA, 2012. ACM.
- [6] Justin Basilico and Yves Raimond. Déjà Vu: The Importance of Time and Causality in Recommender Systems. *Proceedings of the Eleventh ACM Conference on Recommender Systems*, page 342, 2017.
- [7] Veronika Bogina. Addressing Temporal Aspects in User Modelling. *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization - UMAP '17*, pages 113–114, 2017.
- [8] Robin Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370, 2002.
- [9] Robin Burke. The adaptive web. chapter Hybrid Web Recommender Systems, pages 377–408. Springer-Verlag, Berlin, Heidelberg, 2007.
- [10] Pedro G. Campos, Fernando Díez, and Iván Cantador. Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols. *User Modeling and User-Adapted Interaction*, 24(1):67–119, Feb 2014.

- [11] Abhijnan Chakraborty, Saptarshi Ghosh, Niloy Ganguly, and Krishna P. Gummadi. Optimizing the Recency-Relevancy Trade-off in Online News Recommendations. *Proceedings of the 26th International Conference on World Wide Web - WWW '17*, (i):837–846, 2017.
- [12] Chih-Ming Chen, Ming-Feng Tsai, Yu-Ching Lin, and Yi-Hsuan Yang. Query-based Music Recommendations via Preference Embedding. *Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16*, pages 79–82, 2016.
- [13] Chen Cheng, Haiqin Yang, Irwin King, and Michael R. Lyu. Fused matrix factorization with geographical and social influence in location-based social networks. In *AAAI*, 2012.
- [14] Tiago Cunha, Carlos Soares, and Rua Trabalhador Sancarlene. Metalearning for Context-aware Filtering : Selection of Tensor Factorization Algorithms. pages 14–22, 2017.
- [15] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, July 2011.
- [16] Dumitrescu, Doina Alexandra and Simone, Santini. Using Context to Get Novel Recommendation in Internet Message Streams. *WWW 2015 Companion: Proceedings of the 24th International Conference on World Wide Web*, pages 783–786, 2015.
- [17] Daniel M. Fleder and Kartik Hosanagar. Recommender systems and their impact on sales diversity. In *Proceedings of the 8th ACM Conference on Electronic Commerce, EC '07*, pages 192–199, New York, NY, USA, 2007. ACM.
- [18] Huiji Gao, Jiliang Tang, Xia Hu, and Huan Liu. Exploring temporal effects for location recommendation on location-based social networks. In *Proceedings of the 7th ACM Conference on Recommender Systems, RecSys '13*, pages 93–100, New York, NY, USA, 2013. ACM.
- [19] Marko Gasparic. Context-Based IDE Command Recommender System. *Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16*, pages 435–438, 2016.
- [20] Hancheng Ge, James Caverlee, and Haokai Lu. TAPER : A Contextual Tensor-Based Approach for Personalized Expert Recommendation. *10th ACM Conference on Recommender Systems*, pages 261–268, 2016.

- [21] Alexander Tuzhilin Gediminas Adomavicius. *Context-aware recommender systems*, pages 191–226. Springer US, 1 2015.
- [22] Jean-Benoît Griesner, Talel Abdesslem, and Hubert Naacke. POI Recommendation: Towards Fused Matrix Factorization with Geographical and Temporal Influences. In *9th ACM Conference on Recommender Systems, RecSys* , pages 301–304, Vienne, Austria, September 2015. ACM.
- [23] Jungkyu Han and Hayato Yamana. Geographical diversification in poi recommendation: Toward improved coverage on interested areas. In *Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys '17*, pages 224–228, New York, NY, USA, 2017. ACM.
- [24] Daichi Hasumi, Eiji Kamioka, Yassmeen Salman, and et al. Modelling Contextual Information in Session-Aware Recommender Systems with Neural Networks. *Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16*, 42(3):989–992, 2015.
- [25] Daniel Herzog. Recommending a Sequence of Points of Interest to a Group of Users in a Mobile Context. *Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys '17*, pages 402–406, 2017.
- [26] Balázs Hidasi and Domonkos Tikk. Approximate modeling of continuous context in factorization algorithms. *Proceedings of the 4th Workshop on Context-Awareness in Retrieval and Recommendation - CARR '14*, pages 3–9, 2014.
- [27] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*, pages 263–272, Dec 2008.
- [28] Liwei Huang, Yutao Ma, and Yanbo Liu. Point-of-interest recommendation in location-based social networks with personalized geo-social influence. *China Communications*, 12(12):21–31, December 2015.
- [29] Passakorn Khanthaapha, Luepol Pipanmaekaporn, and Suwatchai Kamonsantiroj. Topic-based user profile model for poi recommendations. In *Proceedings of the 2Nd International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence, ISMSI '18*, pages 143–147, New York, NY, USA, 2018. ACM.
- [30] Donghyun Kim, Chanyoung Park, Jinoh Oh, Sungyoung Lee, and Hwanjo Yu. Convolutional Matrix Factorization for Document Context-Aware Recommendation. *Pro-*

- ceedings of the 10th ACM Conference on Recommender Systems - RecSys '16*, pages 233–240, 2016.
- [31] Yehuda Koren. Collaborative filtering with temporal dynamics. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, page 447, 2009.
 - [32] Dominik Kowald. Modeling cognitive processes in social tagging to improve tag recommendations. *Www*, pages 505–509, 2015.
 - [33] Dominik Kowald, Subhash Chandra Pujari, and Elisabeth Lex. Temporal Effects on Hashtag Reuse in Twitter. *Proceedings of the 26th International Conference on World Wide Web - WWW '17*, 5:1401–1410, 2017.
 - [34] Maciej Kula. Metadata embeddings for user and item cold-start recommendations. *CoRR*, abs/1507.08439, 2015.
 - [35] Christopher Laß, Daniel Herzog, and Wolfgang Wörndl. Context-aware tourist trip recommendations. *RecTour 2017*, 2017.
 - [36] Xin Li, Mingming Jiang, Huiting Hong, and Lejian Liao. A time-aware personalized point-of-interest recommendation via high-order tensor factorization. *ACM Trans. Inf. Syst.*, 35(4):31:1–31:23, June 2017.
 - [37] Yanchi Liu, Chuanren Liu, Bin Liu, Meng Qu, and Hui Xiong. Unified point-of-interest recommendation with temporal interval assessment. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1015–1024, New York, NY, USA, 2016. ACM.
 - [38] Andreas Lommatzsch, Benjamin Kille, and Sahin Albayrak. Incorporating context and trends in news recommender systems. In *Proceedings of the International Conference on Web Intelligence*, WI '17, pages 1062–1068, New York, NY, USA, 2017. ACM.
 - [39] Shokirkhon Oppokhonov, Seyoung Park, and Isaac K. E. Ampomah. Current location-based next poi recommendation. In *Proceedings of the International Conference on Web Intelligence*, WI '17, pages 831–836, New York, NY, USA, 2017. ACM.
 - [40] Roberto Pagano, Paolo Cremonesi, Massimo Quadrana, Martha Larson, Balázs Hidasi, Domonkos Tikk, and Alexandros Karatzoglou. The Contextual Turn: from Context-Aware to Context-Driven Recommender Systems. *Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16*, pages 249–252, 2016.

- [41] Róbert Pálovics, András a Benczúr, Levente Kocsis, Tamás Kiss, and Erzsébet Frigó. Exploiting temporal influence in online recommendation. In *Proceedings of the 8th ACM Conference on Recommender systems - RecSys '14*, pages 273–280, New York, New York, USA, 2014. ACM Press.
- [42] Francisco J. Peña. Unsupervised Context-Driven Recommendations Based On User Reviews. *Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys '17*, pages 426–430, 2017.
- [43] Elisa Quintarelli, Emanuele Rabosio, and Letizia Tanca. Recommending New Items to Ephemeral Groups Using Contextual User Influence. *Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16*, pages 285–292, 2016.
- [44] Dimitrios Rafailidis and Alexandros Nanopoulos. Repeat Consumption Recommendation Based on Users Preference Dynamics and Side Information Categories and Subject Descriptors. *WWW 2015 Companion: Proceedings of the 24th International Conference on World Wide Web*, pages 99–100, 2015.
- [45] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 452–461, Arlington, Virginia, United States, 2009. AUAI Press.
- [46] Paul Resnick and Hal R. Varian. Recommender systems. *Commun. ACM*, 40(3):56–58, March 1997.
- [47] Francesco Ricci, Dario Cavada, Nader Mirzadeh, Adriano Venturini, et al. Case-based travel recommendations. *Destination recommendation systems: behavioural foundations and applications*, pages 67–93, 2006.
- [48] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors. *Recommender Systems Handbook*. Springer US, Boston, MA, 2nd editio edition, 2015.
- [49] Markus Schedl, Peter Knees, and Fabien Gouyon. New Paths in Music Recommender Systems Research. *Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys '17*, pages 392–393, 2017.
- [50] Jeffrey C. Schlimmer and Richard H. Granger, Jr. Beyond incremental processing: Tracking concept drift. In *Proceedings of the Fifth AAAI National Conference on Artificial Intelligence*, AAAI'86, pages 502–507. AAAI Press, 1986.

- [51] Ido Tamir, Roy Bass, Guy Kobrinsky, Baruch Brutman, Ronny Lempel, and Yoram Dayagi. Powering Content Discovery through Scalable, Realtime Profiling of Users' Content Preferences. *Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16*, pages 399–400, 2016.
- [52] Alexey Tsymbal. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, 106(2), 2004.
- [53] Denis Vuckovac, Julia Wamsler, Alexander Ilic, and Martin Natter. Getting the Timing Right: Leveraging Category Inter-purchase Times to Improve Recommender Systems. *RecSys '16*, 16:277–280, 2016.
- [54] Weiqing Wang, Hongzhi Yin, Ling Chen, Yizhou Sun, Shazia Sadiq, and Xiaofang Zhou. St-sage: A spatial-temporal sparse additive generative model for spatial item recommendation. *ACM Trans. Intell. Syst. Technol.*, 8(3):48:1–48:25, April 2017.
- [55] Weiqing Wang, Hongzhi Yin, Xingzhong Du, Quoc Viet Hung Nguyen, and Xiaofang Zhou. Tpm: A temporal personalized model for spatial item recommendation. *ACM Trans. Intell. Syst. Technol.*, 9(6):61:1–61:25, November 2018.
- [56] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three, IJCAI'11*, pages 2764–2770. AAAI Press, 2011.
- [57] Jason Weston, Hector Yee, and Ron J. Weiss. Learning to rank recommendations with the k-order statistic loss. In *Proceedings of the 7th ACM Conference on Recommender Systems, RecSys '13*, pages 245–248, New York, NY, USA, 2013. ACM.
- [58] D. Yang, D. Zhang, V. W. Zheng, and Z. Yu. Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1):129–142, Jan 2015.
- [59] Hongzhi Yin, Bin Cui, Ling Chen, Zhiting Hu, and Zi Huang. A temporal context-aware model for user behavior modeling in social media systems. *Proceedings of the 2014 ACM SIGMOD international conference on Management of data - SIGMOD '14*, (1):1543–1554, 2014.
- [60] Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. Time-aware point-of-interest recommendation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13*, pages 363–372, New York, NY, USA, 2013. ACM.

- [61] Matthew D. Zeiler. Adadelata: An adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.
- [62] Jia-Dong Zhang and Chi-Yin Chow. Geosoca: Exploiting geographical, social and categorical correlations for point-of-interest recommendations. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 443–452, New York, NY, USA, 2015. ACM.
- [63] Shenglin Zhao. *Point-of-interest Recommendation in Location-based Social Networks*. PhD thesis, 2018.
- [64] Shenglin Zhao, Tong Zhao, Irwin King, and Michael R. Lyu. Geo-teaser: Geo-temporal sequential embedding rank for point-of-interest recommendation. In *WWW*, 2017.
- [65] Yong Zheng, Mayur Agnani, and Mili Singh. Identifying grey sheep users by the distribution of user similarities in collaborative filtering. In *Proceedings of the 6th Annual Conference on Research in Information Technology*, RIIT '17, pages 1–6, New York, NY, USA, 2017. ACM.

A Plán práce na riešení projektu

Plán priebehu prác na správe o projekte - Diplomový projekt I:

- I. Týždeň - iniciálny prieskum ohľadom témy vrámci vedeckých článkov a taktiež sumarizácia pokynov k vypracovaniu správy,
- II. Týždeň - čítanie článkov,
- III. Týždeň - čítanie článkov,
- IV. Týždeň - sumarizácia prvotných článkov a hľadanie datasetu,
- V. Týždeň - pokračovanie v čítaní článkov a hľadaní datasetu,
- VI. Týždeň - pokračovanie v čítaní článkov a hľadaní datasetu,
- VII. Týždeň - zameranie sa na konkrétnu vybranú doménu, začiatok spisovania správy,
- VIII. Týždeň - pokračovanie písania správy i čítania článkov v už konkrétnej oblasti,
- IX. Týždeň - pokračovanie písania správy i čítania článkov v už konkrétnej oblasti,
- X. Týždeň - písanie správy a finálne potvrdenie datasetu,
- XI. Týždeň - písanie správy,
- XII. Týždeň - finálne úpravy na odovzdanie.

V realite nešlo všetko úplne podľa predpokladaných krokov - problém bol najmä vybrať ten správnu dátovú sadu a doménu. Dlhو sme zvažovali doménu masmédií, konkrétne novinových článkov, čo sa ale nakoniec neukázalo ako perspektívne, hoci sa pôvodne črtalo niekoľko smerov, ktorými by sa dalo ísť. V tejto fáze sa plánovalo navrhnuť metódu schopnú odolávať problému zmeny konceptu.

Kvôli tomuto zdržaniu sa potom práce posunuli o niekoľko týždňov, resp. dopísanie práce až na predposledný týždeň, keďže finálny návrh je diametrálne iný od pôvodne zamýšľaného.

Predpokladaný plán práce na projekte v druhom semestri - Diplomový projekt II - bol nasledovný:

- I. Týždeň - zisk dát,
- II. Týždeň - finalizácia zisku dát, začiatok ich analýzy,

- III. Týždeň - pokračovanie a finalizácia analýzy dát,
- IV. Týždeň - vývoj základného prototypu odporúčacieho systému,
- V. Týždeň - vývoj základného prototypu odporúčacieho systému na základe kontextu,
- VI. Týždeň - vývoj základného prototypu odporúčacieho systému, ktorý berie v úvahu temporálne aspekty,
- VII. Týždeň - porovnávanie metód a ich postupné vylepšovanie, písanie správy,
- VIII. Týždeň - finálne úpravy programu, postupné spisovanie správy,
- IX. Týždeň - overovanie metódy, hlavná fáza spisovania konečnej správy,
- X. Týždeň - vyhodnocovanie metódy, pokračovanie v písaní finálnej správy,
- XI. Týždeň - finalizácia písomnej časti práce,
- XII. Týždeň - drobné úpravy na písanej práci, odovzdanie.

Harmonogram nebol presne dodržaný - kľúčovým problémom boli komplikácie spojené s účasťou na zahraničnej mobilite Erasmus (absencia stabilného bývania a internetu prvý mesiac, nečakaná zmena syllabov na hosťujúcej univerzite a s tým spojené ďalšie starosti). Napriek tomu sa podarilo nabehnúť na plán približne v III. týždni, kedy sa podarilo dosiahnuť prvé výsledky, ktoré sa následne iteratívne zlepšovali a skonkrétňovali. Veľká zmena nastala v prístupe, rozhodli sme sa ísť cestou iteratívneho prístupu miesto pôvodného plánu, ktorý pripomína skôr vodopád.

Predpokladaný plán riešenia Diplomového projektu III. bol nasledovný:

- I. Týždeň - zapracovanie geografického vplyvu,
- II. Týždeň - zapracovanie temporálneho vplyvu, vypracovanie článku na IIT.SRC,
- IV. Týždeň - optimalizovanie výsledkov,
- V. Týždeň - optimalizovanie výsledkov,
- VI. Týždeň - otestovanie nad iným datasetom, optimalizácia, písanie správy, (približne v tomto období: finálne úpravy na článku IIT.SRC)
- VII. Týždeň - otestovanie nad iným datasetom, optimalizácia, písanie správy,
- VIII. Týždeň - finálne úpravy programu, písanie správy, príprava na IIT.SRC,

- IX. Týždeň - hlavná fáza spisovania konečnej správy,
- X. Týždeň - drobné úpravy na písanej práci, prezentácia na IIT.SRC,
- XI. Týždeň - odovzdanie.

Ciele plánu sa podarili splniť, i keď nie v rovnakom poradí. Časť sa vykonala ešte pred začiatkom semestra, pričom ale veľa času bolo stráveného nad implementáciou modelovania. Kvôli príliš stručnej dokumentácii knižnice LightFM sa nám jednu chybu nedarilo odstrániť niekoľko mesiacov. Následne sa zdržal postup aj z dôvodu kolízie semestrov na domovskej a Erasmus+ univerzite – zimný semester končil až po začiatku letného semestra, posledná skúška zimného semestra bola vykonaná približne sedem týždňov po začiatku letného semestra (prvý riadny termín). Preto bolo vykonávanie experimentov na začiatku semestra pomalšie.

Plán sa už do konca semestra nepodarilo dobehnúť, ale najmä z dôvodu, že z pôvodne zamýšľaného riešenie sa nakoniec stali dva rôzne prístupy, ktoré samozrejme vyžadovali takmer dvojnásobok času a úsilia.

B Prehľad aktuálnych riešení temporálnych odporúčacích systémov

Vyhotovený prehľad aktuálnych riešení odporúčacích systémov založených na temporálnych aspektoch (Tabuľka č. 15). Práce sa zameriavajú na rôzne domény.

Tabuľka 15: *Prehľad aktuálnych riešení odporúčacích systémov zohľadňujúcich temporálne aspekty.*

Autor	Názov	Dataset	Metóda	Metrika	Zdroj
Anelli, Vito W. et al.	An Analysis on Time- and Session-aware Diversification in Recommender Systems	Netflix Prize Context	temporálne varianty Explicit Query Aspect Diversification (xQuAD, rámec pre diverzifikáciu)	NDCG, ERR-IA, vlastné metriky	[4]
Lommatzsch, Andreas et al.	Incorporating context and trends in news recommender systems	CLEF News-REEL challenge	porovnávanie odporúčacích modelov - na základe periodicity alebo novosti položiek	kliknutia, miera pre-kliku	[38]
Pálovics, Róbert et al.	Exploiting temporal influence in online recommendation	Last.fm (70k používateľov)	faktORIZÁCIA MATÍC	presnosť, úplnosť, NDCG	[41]
Schedl, Markus et al.	New Paths in Music Recommender Systems Research	LFM-1b dataset	-tutoriál-	-tutoriál-	[49]
Chen, Chih-Ming et al.	Query-based Music Recommendations via Preference Embedding	LFM-1K, Million Song Dataset, KKBOX Inc.	zakódovanie používateľských preferencií do nízko-dimenzionálnych vektorových priestorov	presnosť@k, mAP@k	[12]
Hidasi, Balázs Tikk, Domonkos	Approximate modeling of continuous content in factorization algorithms	Last.fm-1K, TV1, TV2, Grocery, VoD	fuzzy kontextuálne modelovanie, ako kontext - sezónnosť, zahrnuté do algoritmu iTALS (faktORIZÁCIA)	úplnosť@20	[26]
Tamir, Ido et al.	Powering Content Discovery through Scalable, Realtime Profiling of Users' Content Preferences	Outbrain	personalizácie sú kombinované s odporúčaním založenom na obsahu a kolaboratívnym odporúčaním		[51]
Gasparic, Marko	Context-Based IDE Command Recommender System	Eclipse IDE	ešte len plánujú zahrnúť zhromaždené temporálne aspekty	metriky správnosti	[19]
Kim, Donghyun et al.	Convolutional Matrix Factorization for Document Context-Aware Recommendation	Movie-Lens, Amazon + doplnené z IMDB	konvolučná faktORIZÁCIA MATÍC (spojenie konvulčnej neurónovej siete a pravdepodobnostnej MF)	RMSE	[30]

Yin, Hongzhi et al.	A temporal context-aware model for user behavior modeling in social media systems	Digg, Lens, Movie, Delicious	model latentnej triedy štatistických kombinácií, konkrétne model časového kontextuálneho kontaktu	presnosť@k, NDCG@k	[59]
Koren, Yehuda	Collaborative filtering with temporal dynamics	Netflix	modelovanie temporálnej dynamiky počas celého obdobia kolaboratívnymi technikami (MF)	RMSE	[31]
D.A. Dumitrescu, S. Simone	Using Context to Get Novel Recommendation in Internet Message Streams	Reuters Volume 1	kontextuálne filtrovanie pomocou algoritmu na základe novosti	mAP, stredné skóre novosti	[16]
Dominik Kowald	Modeling cognitive processes in social tagging to improve tag recommendations	BibSonomy, CiteULike, Delicious	kognitívny model ACT-R	presnosť@k, úplnosť@k	[32]
Rafailidis Dimitrios, Nanopoulos Alexandros	Repeat Consumption Recommendation Based on Users Preference Dynamics and Side Information Categories and Subject Descriptors	Last.fm, MovieLens	CTF a W-CTF s modelovaním (používateľské preferencie a vedľajšie informácie, znižovanie váh preferencií)	správnosť a TF	[44]
Abhijnan Chakraborty et al.	Optimizing the Recency-Relevancy Trade-off in Online News Recommendations	Yahoo! News Dataset, CLEF News Dataset, NYTimes dataset	optimalizácia rámca kombináciou predikovaného budúceho dopadu príbehov s neistotami v predikcii	správnosť	[11]
Dominik Kowald et al.	Temporal Effects on Hashtag Reuse in Twitter: A Cognitive-Inspired Hashtag Recommendation Approach	Twitter	ACT-R (najprv v kombinácii s temporálnymi vzormi správania a potom aj s odporúčaním na základe obsahu)	presnosť, úplnosť, F1, MRR, mAP, NDCG	[33]
Kowald, Dominik	Modeling cognitive processes in social tagging to improve tag recommendations	34 datasetov (Amazon Reviews, TripAdvisor, Yelp...)	výber tenzor faktorizácie pre filtrovanie založené na obsahu	NDCG, stredná priemerná presnosť, MRR	[14]

C Ukážka implementácie odporúčania s knižnicou LightFM

Vytvorenie matice používateľa U x položky I pre tréning (podobne aj pre testovanie) z matice hodnotení U x I :

```
from scipy.sparse import csr_matrix

user_u = list(sorted(ratings.user_id.unique()))
business_u = list(sorted(ratings.business_id.unique()))

data = ratings['rating'].tolist()
row = ratings.user_id.astype('category',
                             categories=user_u).cat.codes
col = ratings.business_id.astype('category',
                                 categories=business_u).cat.codes

sparse_matrix = csr_matrix((data, (row, col)),
                           shape=(len(user_u), len(business_u)))
```

Inicializácia najdôležitejšieho komponentu pre LightFM odporúčanie – modelu. Modelu sa môže zadať viacero parametrov, napríklad stratová funkcia. Na konci ho používame v kombinácii s predtým vytvorenou maticou:

```
from lightfm import LightFM
model = LightFM(loss='warp', learning_schedule='adadelta',
                no_components=30)
model.fit(sparse_matrix, epochs=20)
```

Generácia odporúčaní pre daného používateľa:

```
predictions = model.predict(user_id, np.arange(len(item_ids)))
```

Evaluácia výsledkov implementovaná vrámci LightFM – presnosť@k, k=3:

```
from lightfm.evaluation import precision_at_k
precision_at_k(model, sparse_matrix_test, k=3).mean()
```

Evaluácia výsledkov, pokrytie:

```
from lightfm.evaluation import recall_at_k  
recall_at_k(model, sparse_matrix_test, k=3).mean())
```

Funkcie evaluácie sme v práci nevyužili pri generácii hlavných výsledkov. Miesto toho sme si implementovali vlastnú upravenú metódu.

D Tabuľky výsledkov pre jednotlivé črty modelovania na dátovej sade B

Táto príloha prezentuje bližšie výsledky jednotlivých črt pri modelovaní. Tabuľka č. 16 zobrazuje výsledky pre črty sezónnosti, kde VP znamená vážený priemer, AP aritmetický priemer, 3M pre tri mesiace. Je prirodzené, že výsledky pre štát sú rovnaké ako pre základné odporúčanie (mierny rozdiel, vstup náhody). Z črt kategórie dosahuje najlepšie výsledky vážený priemer kategórií (pre aktuálny a dva susedné mesiace). Výsledky medzi rôznou formou mesiaca sú minimálne, avšak bol zvolený cyklický, keďže lepšie vystihuje charakter mesiacov v roku.

Tabuľka 16: Porovnanie črt pre modelovanie na základe sezónnosti.

Črta	Kategórie one hot		VP		VP 3M	
k	Presnosť	Pokrytie	Presnosť	Pokrytie	Presnosť	Pokrytie
1	1,04	0,44	1,17	0,59	1,65	0,93
3	1,3	1,13	1,54	1,39	1,93	1,73
5	1,81	1,75	1,83	1,81	2,36	2,31
10	3,71	3,69	3,52	3,51	3,95	3,95
Črta	Štát		AP		AP 3M	
1	1,21	0,55	1,49	0,58	1,23	0,45
3	1,70	1,53	1,83	1,68	1,71	1,55
5	2,62	2,55	2,35	2,33	2,24	2,21
10	3,69	3,68	3,61	3,61	3,62	3,62
Pred-F,črta	-		Mesiac		Mesiac cyklicky	
1	1,4	1,21	1,39	1,2	1,14	1,1
3	2,7	2,7	2,7	2,7	2,9	2,9
5	3,33	3,33	3,5	3,5	3,57	3,57
10	3,72	3,72	4,23	4,23	4,19	4,19

Tabuľka č. 17 zobrazuje výsledky pre dlhodobé trendy, kde VP znamená vážený priemer, AP aritmetický priemer. Pokúsili sme sa využiť ako črtu aj najlepšiu metódu skóre pre po-filter, ale neúspešne. Aritmetický priemer rozdielov získal obrovský náskok pred ostatnými metódami.

Podrobný opis črt sa nachádza v návrhu, Kapitole č. 5.

Tabuľka 17: Porovnanie črt pre modelovanie na základe dlhodobého trendu.

Črta	Kategórie one hot		VP Rozdiel		AP Rozdiel	
k	Presnosť	Pokrytie	Presnosť	Pokrytie	Presnosť	Pokrytie
1	3,18	0,85	3,67	1,07	4,56	1,02
3	3,5	2,13	3,34	2,19	4,16	2,52
5	3,84	3,18	3,35	2,84	4,39	3,52
10	5,95	5,67	4,74	4,58	7,58	7,22
Črta	Štát		VP Rozdiel %		AP Rozdiel %	
1	2,49	0,68	2,42	5,1	2,7	0,74
3	3,42	2,19	2,93	1,94	3,4	2,27
5	4,13	3,47	3,53	2,97	3,69	3,23
10	5,45	5,24	5,21	5	5,29	5,11
Črta	Rozdiel norm.arit.(ako po-F), štát		VP Rozdiel norm.		AP Rozdiel norm.	
1	2,62	5,53	3,18	0,88	3,18	0,88
3	2,84	1,71	3,12	2,05	3,12	2,05
5	3,57	2,98	3,65	3,06	3,65	3,06
10	5,07	4,87	5,67	5,46	5,67	5,46

E Článok prezentovaný na študentskej vedeckej konferencii IIT.SRC 2019

Finálna verzia článku prezentovaného na študentskej vedeckej konferencii Fakulty informatiky a informačných technológií, Slovenskej technickej univerzity v Bratislave. Názov konferencie je IIT.SRC 2019.

Temporal Recommendations based on Locality-specific Seasonality and Long-term Trends

Elena ŠTEFANCOVÁ*

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

Abstract. This work deals with time-aware recommender systems in a domain of location-based social networks, such as Yelp or Foursquare. We propose a novel method to recommend Points-of-Interest (POIs) which consider their seasonality and long-term trends. In contrast to existing methods, we model these temporal aspects specifically for individual geographical areas instead of globally. In addition, a geographical post-filter method is used for creating personal regions of users. The preliminary results show that consideration of locality-specific seasonality and long-term trends in categories' popularity can improve the performance of the proposed recommender system.

1 Introduction

Recommender systems are an important part of web services and their popularity is on increase. They are used by e-shops, streaming services, newspaper portals etc. to keep and satisfy the customer. Taking context (e.g. time, location) into account in some cases was proven to improve the accuracy of recommendations. In the case of temporal context, recommendation can reflect changes of trends in society and cyclic patterns in the behaviour of people – for example, a user buys different clothes during a summer and during a winter, goes for a coffee in the morning and to a bar in the evening.

We focus on the domain of location-based social networks, which provide users with possibility to search and rate so called Points-of-Interest (POIs). In the context of POIs, recommendation plays an important role (e.g. to suggest users which shops or restaurants may be interesting for them). This kind of social networks is well known due to the importance of geographical and temporal context for recommendation. We focus specifically on temporal recommenda-

tions which take seasonality and long-term trends into account. Some previous works have already focused on seasonality and trends within POI recommendation, but usually only on a global level and they have not considered locality specifics (i.e. seasonality and trends specific for a particular geographic area influenced by local climate). We decided to address this open problem in our work.

By means of data pre-filtering and modelling with a matrix factorization, we incorporated the mentioned temporal aspects into POI recommendation. For improving the results, we used a geography filtering, too. This is based on so called “personal user regions”, what is taking into account which areas have been already visited by the user.

Our main contribution is a research of the dependence of the temporal aspects on the locality.

2 Context in Recommender Systems

Many approaches of recommending exists – collaborative filtering, content-based filtering, demographic, knowledge-based, hybrid [3] – but all of them have in common that the final recommendation results from function of users and items (Equation n. 1, users as U , items as I) [1,3].

$$R : U \times I \rightarrow \text{Rating} \quad (1)$$

Within the classical recommendation techniques any context is not taken into account. However, in many cases its inclusion is relevant – users watch different movies during Christmas and during the rest of the year, parents choose different movies for them and for their children etc. [1]. The most popular aspects of the context are time, locality or company of other people. For the context recommendation, the ratings

* 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

are not only function of users and times, but also of the context (Equation n. 2, U for users, I for items, C for context) [1].

$$R : U \times I \times C \rightarrow Rating \quad (2)$$

The incorporation of the context can be done by [2, 3]:

- Pre-filtering – a segment of ratings is filtered according to relevance of the context, then selected part is used for recommendation process.
- Post-filtering – first the ratings are generated, then reordered based on the context.
- Modelling – information of context are used during the whole process of recommendation.

Temporal aspects of items and users can be based on recency (the recent transactions are the most relevant) or periodicity (change of behaviour based on time during the date, day of the week etc.) [3, 5].

But any type of contextual recommendation is still prone to the typical recommender systems problems, e.g. long tail (problematic distribution of ratings for collaborative approaches – few popular items cover most of the transactions, most of the items is rated rather rarely) [3].

2.1 Time-aware POI Recommendation

The social networks have been never so popular as today. One type of them are Location-Based Social Networks¹ (LBSN) [4], e.g. Foursquare², Yelp³ or Gowalla⁴. The basic element of LBSN is a Point-of-Interest (POI). Check-ins serve as a transactions in LBSN (corresponding to POI visits of a particular user at the given time) [7, 8, 10, 11, 13] or reviews of POIs [6, 12]. While users usually create check-ins for the same POI repetitively, reviews are created only once typically after the first POI visit.

Since LBSNs try to motivate users to visit a lot of new localities and share the experience, it is crucial to make the recommendation as relevant as possible and context aspects such as locality (e.g. a user is willing to visit POIs only in some areas or in a certain distance from his/her usual places of occurrence) and time (e.g. preferences of users are strongly cyclic during the day/week/year) are necessary [7, 11, 12]. Influence of POI categories is significant, too. Every user has a personal taste.

For employing the locality so called home locality (home address or the most visited area) [10] or Personal Functional Regions (several visited areas, sometimes connected to particular type of activities) [11] are used.

Temporal aspects are often used with combination of collaborative filtering, since people tend to have similar cyclic patterns [7, 11]. For time-aware recommendation is a typical problem of data sparsity – low frequency of check-ins [7, 12] resulting from the explicit nature of the data. In case of huge sparsity of the data, methods revealing latent properties are popular. Such a state-of-the-art method of recommendation is matrix factorization [9].

The most used evaluation methods are precision@k and recall@k [8, 11, 13].

3 Method proposal

We propose a method of time-aware recommendation to explore seasonality and long-term trends of POI categories' popularity based on geographical locality (Figure 1). In contrast to many existing methods, we aim to recommend the first visit of so far not visited POI (represented by reviews) instead of repetitive visits (represented by check-ins). A reason for this decision is a number of missing check-ins since users many times skip check-in and thus check-in data may not reflect actual POI visits. The method combines temporal pre-filtering, modelling of temporal aspects and geographical post-filtering.

The main recommendation method is based on combination of collaborative filtering (basic matrix Users x POIs) and matrix factorization. The given ratings are normalized for every user on his own to positive/negative. It is possible by using item features (e.g. category) to employ a content-based approach, too.

The first step is a temporal pre-filtering which is used mainly for the seasonality. Transactions of every POI are divided according to a month and because of that, one original item (POI) can result into up to 12 new items (month instances of the original POI, every one of them has in its features encoded the original POI).

After the pre-filter, we use a modelling method (based on combination of content-based recommendation and collaborative filtering) with matrix factorization. Item feature vectors are used as a storage of context and item information. The features for the sea-

¹ https://en.wikipedia.org/wiki/Geosocial_networking

² <https://foursquare.com/>

³ <https://www.yelp.com>

⁴ <https://en.wikipedia.org/wiki/Gowalla>

sonality of the categories: month as a cyclic ordinal value in a form of sinus and cosinus value, the state as one-hot coding, category in a form of one-hot coding. For the long term trends in categories: average exponential factor of rating decay for all business of a category, average change in the number of visits in comparison with the previous season (year/month) for all business of a category, the state as one-hot coding, category in a form of one-hot coding.

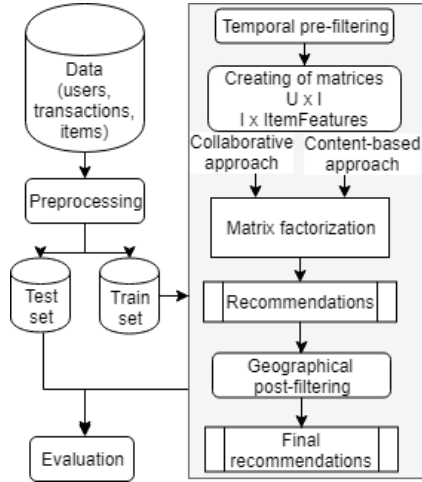


Figure 1. Scheme of the proposed method.

The used loss function is WARP – it is suitable for optimization of the highest items on the recommendation list. As a learning rate schedule adadelata is used – it gives better results in combination with low-number-of-epochs WARP, so for the less time cost version.

After generating of the recommendations, we reorder them based on their locality (geographical post-filtering) – the ones within the certain distance (threshold of 0.5 km) from previously visited areas are preferred before the rest of them.

4 Experimental evaluation

We chose Yelp dataset⁵ for the method verification. We used two subsets of it – North Las Vegas data (1 269 users, 1 142 POIs and 11 700 reviews) and Markham data (about the same number of users/POIs/reviews). Only users with 5 and more reviews are included and categories with less than 100 businesses are merged with a category higher in the hierarchy. Every business has at least 3 reviews. A review include a 1-5 stars rating. The data were divided to

train and test sets according to the time in ratio 80:20. The train ratings are normalized to 1/0/-1 values for every used based on his average ratings. The context-aware recommendation was performed with a use of LighFM⁶ library.

4.1 Preliminary results

First, we compared results of basic collaborative recommendation with matrix factorization and its combination with geographic post-filtering (Table 1). As expected, the post-filtering helped to improve recommendation performance significantly.

Table 1. Collaborative recommendation with matrix factorization and extension of geo. post-filtering.

Method	P@3	R@3
Dataset	Test	Test
MF	0,70	0,29
MF with geo.	0,87	0,49

Then we compared results for basic collaborative recommendation with matrix factorization and its pre-filtered version (Table 2). Again, the improvement of results is visible. This claim supports also the exploratory analysis – we can observe stronger bias of a reviews sum for summer and winter months in cities with a larger gap of the temperatures during the year (Figure 2).

Table 2. Collaborative recommendation with matrix factorization in month intervals before and after extension of contextual aspects.

Dataset	N. Las Vegas		Markham	
Method	P@3	R@3	P@3	R@3
Dataset	Test	Test	Test	Test
MF	0,10	0,22	0,04	0,07
MF+pre-filt.	0,11	0,26	0,05	0,09

The final step was adding item features to the temporal pre-filtering version (last version, from Table 2). It is surprising that the values drop after involving of the features – this can be, however, explained as a possible implementation issue, which we are going to address in our future work (Table 3). However, combination of month and category seems promising, but we have to figure out how to code the features properly, without decreasing the precision and recall.

⁵ <https://www.yelp.com/dataset>

⁶ <https://github.com/lyst/lightfm>

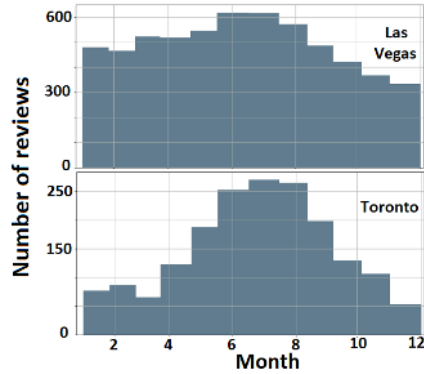


Figure 2. Number of reviews in a city with weak(Las Vegas)/strong(Toronto) winters divided according to months.

Table 3. Matrix factorization, temporal pre-filtering and POI features ($1M = \text{identity matrix} - I \times I$).

Method	P@3	R@3
Dataset	Test	Test
Without features	0,110	0,260
(+1M)	0,003	0,008
Month (+1M)	0,005	0,006
Category (+1M)	0,002	0,003
Month, category (+1M)	0,013	0,025

5 Conclusion

We researched temporal aspects in the domain of LBSN – how the including seasonality and long-term trends in popularity of POI categories can affect the results of a recommender system. We design a method employing both geographical and temporal aspects and evaluated it on Yelp dataset. The first results are promising - both temporal and geographical aspects improved precision@k and recall@k.

The next steps consist mostly of correcting the implementation of the feature vectors. After that adding and optimizing of the features will be done.

References

- [1] Adomavicius, G., Sankaranarayanan, R., Sen, S., Tuzhilin, A.: Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach. *ACM Trans. Inf. Syst.*, 2005, vol. 23, no. 1, pp. 103–145.
- [2] Adomavicius, G., Tuzhilin, A.: Context-aware Recommender Systems. *Proceedings of the 2008 ACM Conference on Recommender Systems*, 2008, pp. 335–336.
- [3] Aggarwal, C.C.: *Recommender Systems - The Textbook*. Volume 40. Springer International Publishing, Cham, 2016.
- [4] Bao, J., Zheng, Y., Mokbel, M.F.: Location-based and Preference-aware Recommendation Using Sparse Geo-social Networking Data. In: *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*. SIGSPATIAL '12, New York, NY, USA, ACM, 2012, pp. 199–208.
- [5] Basilico, J., Raimond, Y.: Déjà Vu: The Importance of Time and Causality in Recommender Systems. *Proceedings of the Eleventh ACM Conference on Recommender Systems*, 2017, p. 342.
- [6] Han, J., Yamana, H.: Geographical Diversification in POI Recommendation: Toward Improved Coverage on Interested Areas. In: *Proceedings of the Eleventh ACM Conference on Recommender Systems*. RecSys '17, New York, NY, USA, ACM, 2017, pp. 224–228.
- [7] Li, X., Jiang, M., Hong, H., Liao, L.: A Time-Aware Personalized Point-of-Interest Recommendation via High-Order Tensor Factorization. *ACM Trans. Inf. Syst.*, 2017, vol. 35, no. 4, pp. 31:1–31:23.
- [8] Oppokhonov, S., Park, S., Ampomah, I.K.E.: Current Location-based Next POI Recommendation. In: *Proceedings of the International Conference on Web Intelligence*. WI '17, New York, NY, USA, ACM, 2017, pp. 831–836.
- [9] Peña, F.J.: Unsupervised Context-Driven Recommendations Based On User Reviews. *Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys '17*, 2017, pp. 426–430.
- [10] Wang, W., Yin, H., Chen, L., Sun, Y., Sadiq, S., Zhou, X.: ST-SAGE: A Spatial-Temporal Sparse Additive Generative Model for Spatial Item Recommendation. *ACM Trans. Intell. Syst. Technol.*, 2017, vol. 8, no. 3, pp. 48:1–48:25.
- [11] Yang, D., Zhang, D., Zheng, V.W., Yu, Z.: Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2015, vol. 45, no. 1, pp. 129–142.
- [12] Zhang, J.D., Chow, C.Y.: GeoSoCa: Exploiting Geographical, Social and Categorical Correlations for Point-of-Interest Recommendations. In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '15, New York, NY, USA, ACM, 2015, pp. 443–452.
- [13] Zhao, S., Zhao, T., King, I., Lyu, M.R.: Geo-Teaser: Geo-Temporal Sequential Embedding Rank for Point-of-interest Recommendation. In: *WWW*, 2017.

F Technická dokumentácia

Pre prevedenie navrhnutých metód sme vytvorili sériu súborov pomocou webovej aplikácie Jupyter Notebook v jazyky Python, verzia 3.6. Podrobné informácie o inštalácii je možné nájsť na webstránke <http://jupyter.org/install.html>, odporúča sa však použiť distribúcia Anaconda, <https://www.continuum.io/downloads>, ktorá umožňuje súčasnú inštaláciu Python aj Jupyter Notebook.

Použité knižnice:

- glob (načítanie dátových sád),
- Matplotlib (vykreslenie grafov),
- pandas (analýza a spracovanie dát),
- NumPy (numerické operácie s polami),
- scipy (práca s maticami),
- seaborn (štatistická dátová vizualizácia),
- LightFM (odporúčanie).

Väčšina notebookov je jednoducho znovupoužiteľných pre iné dátové sady, stačí v úvodných bunkách nastaviť pár premenných (v notebookoch plniacich rôzny účel je aj kombinácia týchto premenných rôzna):

- `f_path` (cesta k dátovej sade),
- `f_path_subset` (cesta ku doplnkovým údajom danej dátovej sady),
- `name_subset` (názov dátovej sady),
- `method_sub` (názov metódy – netreba meniť ak sa notebook neupravil),
- `f_results_path` (priečinkok, kam sa uložia výsledky),
- `epochs_my` (vhodná hodnota parametra `epochs` pre odporúčanie LightFM – volí sa pre každú dátovú sadu zvlášť),
- `no_components_my` (vhodná hodnota parametra `no_components` pre odporúčanie LightFM – volí sa pre každú dátovú sadu zvlášť),
- `learning_schedule_my` (vhodná hodnota parametra `learning_schedule` pre odporúčanie LightFM).

Vstupná dátová sada by mala byť pritom vo formáte CSV, používateľov označených ako `user_id` a `POI` ako `business_id`. Naše riešenie pozostáva primárne z priečinkov s dátami a súborov typu Jupyter Notebook. Všetko je očíslované a zoradené podľa postupu riešenia. Chýbajú akurát pôvodné dáta z Yelpu a niektoré podmnožiny, keďže ich veľkosť je niekoľko GB. Poskytnuté sú najviac používané podmnožiny. V prípade záujmu o celý dataset, jeho aktuálnu verziu je možné stiahnuť z www.yelp.com/dataset a konvertovať z JSON súborov na CSV pomocou priloženého skriptu.

G Opis digitálnej časti práce

Evidenčné číslo práce v informačnom systéme: FIIT-182905-55695

Obsah digitálnej časti práce (archív ZIP):

- Priečinok *Implementácia a dáta*:

- Priečinky:

- * *analysis* - notebooky s kompletnou exploratívnou analýzou pre podniky, prihlásenia, recenzie, tipy, používateľov a temporálnych vplyvov medzi nimi,
- * *data_00_csv* - miesto na uskladnenie kompletnej Yelp dátovej sady, prázdny kvôli príliš veľkému objemu dát,
- * *data_01_processed* - spracovaná dátová sada Yelp na len tie informácie, ktoré sú potrebné, prázdny kvôli príliš veľkému objemu dát,
- * *data_01_processed_subsets* - podmnožiny z *data_01_processed* a tie ďalej spracované na vstupné dáta do po-filtrovaní alebo modelovania, súbory CSV pre jednotlivé dátové sady,
- * *data_01_sample_DATASET* - vybraná menšia dátová sada z celkovej dátovej sady Yelp, CSV súbory,
- * *data_02_sample_months_DATASET* - trénovacia a testovacia množina rozdelená podľa mesiacov posledného roka dát, CSV súbory, pre kombinácie XYZW prázdne kvôli príliš veľkému objemu dát,
- * *results* - výsledky experimentov,

- Jupyter notebooky:

- * *00 Basic Matrix factorization with LightFM.ipynb* - ukážka použitia odporúčania pomocou LightFM,
- * *00 Data for the method.ipynb* - generovanie potrebných vstupov pre metódy pre jednotlivé dátové sady,
- * *00 Sample dataset DATASET - 00.ipynb* - vygenerovanie istej podmnožiny z celkovej dátovej sady Yelp,
- * *00 Sample dataset DATASET - 01 months.ipynb* - rozdelenie dátovej množiny na 12 trénovacích a testovacích množín podľa mesiacov,
- * *01 Basic recommendation - 00 - basic.ipynb* - základné odporúčanie na podmnožine dát,
- * *01 Basic recommendation - 01 - geographical post filter.ipynb* - experimenty s geografickým po-filtrom,

- * prefix *11 Method2 - Options - Dataset DATASET* - experimenty so sezónnosťou pomocou po-filtrovaní,
- * prefix *12 Method2* - experimenty s dlhodobými trendmi pomocou po-filtrovaní,
- * prefix *21 Method* - experimenty so sezónnosťou pomocou modelovania,
- * prefix *22 Method* - experimenty s dlhodobými trendmi pomocou modelovania,
- * prefix *31* - overenie sezónnosti:
 - *00 DATASET.ipynb* - základné odporúčanie,
 - *- 00 - DATASET 00 state.ipynb* - lokálna sezónnosť,
 - *- 00 - DATASET 01 non state.ipynb* - globálna sezónnosť,
- * prefix *32* - overenie dlhodobých trendov:
 - *00 DATASET.ipynb* - základné odporúčanie,
 - *- 00 - DATASET 00 state.ipynb* - lokálne trendy,
 - *- 00 - DATASET 01 non state.ipynb* - globálne trendy,
- * *Avg results.ipynb* - rátať priemeru výsledkov z viacerých behov na základe udaných parametrov,
- * *json_to_csv_converter.py* - Python skript na konvertovanie Yelp JSON súborov na CSV.

- *Anotácie.pdf* - anotácie,
- *DP_ElenaŠtefancová.pdf* - finálna diplomová práca.

Vysvetlivky:

DATASET - označenie dátovej sady (avšak notebook je aplikovateľný po vnútornej zmene parametra na ktorúkoľvek inú)

Digitálna časť práce má veľkosť 2,8 GB, kvôli čomu je uložená v systéme G Suite for Education. Názov odovzdaného archívu: DP_prilohy_digital_ElenaŠtefancová.zip