Včasná identifikácia trendov v správaní používateľov elektronického zľavového portálu

Ondrej Kaššák, Mária Bieliková

Ústav informatiky, informačných systémov a softvérového inžinierstva,
Fakulta informatiky a informačných technológií, Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava

ondrej.kassak@stuba.sk, maria.bielikova@stuba.sk

Abstrakt. Správanie používateľov služieb na webe sa v čase mení. Napríklad v zľavovom portáli používatelia reagujú na jednotlivé ponuky rozlične. Našim cieľom je dokázať včas identifikovať, ktoré ponuky sa stanú trendami (vysoko predávanými) a ktorým naopak treba pomôcť napríklad dodatočnou propagáciou. Hlavnou výzvou tejto úlohy je veľké množstvo dát o nákupoch, ktoré prichádzajú formou kontinuálneho prúdu. Riešenie, ktoré v práci navrhujeme je založené na časovo a výpočtovo efektívnom jednoprechodovom spracovaní dát umožňujúcom prácu v online čase. Týmto spôsobom sme schopní pomerne skoro identifikovať, ktoré zľavy sa stanú trendami. Opisované riešenie sme overili na reálnej množine dát zľavového portálu, kde sme ukázali, že časť trendov je možné identifikovať už na základe prvých dní, kedy sú dané zľavy v ponuke.

Kľúčové slová: frekventované prvky, posuvné okno, prúd dát, trendy v správaní používateľov webovej služby

1 Úvod

Predmetom nášho záujmu je pojem trend. Pod týmto pojmom v doméne online zľavového portálu rozumieme položku nakupovanú používateľmi v určitom časovom období (prípadne celkovo) viac ako iné položky. V oblasti dolovania dát trend typicky predstavuje frekventovaný vzor, teda sekvenciu (asociačné pravidlo), prvok alebo štruktúru (strom alebo graf) [1]. Pre potreby našej úlohy, ktorou je včasná identifikácia nakupovaných položiek, sme vybrali reprezentáciu prostredníctvom frekventovaných prvkov. Tie poskytujú jednoduchú reprezentáciu a zároveň umožňujú identifikáciu trendov v doménach kde používatelia typicky nakupujú jedinú položku (napr. zľavový portál).

Trendom môže byť *top-n* najkupovanejších položiek prípadne položky, ktorých nákup tvorí viac ako *m%* všetkých realizovaných nákupov za sledované obdobie [1]. Keďže *top-n* prvkov nedokáže dynamicky reflektovať počet aktuálne existujúcich trendov, ktorý sa v čase môže meniť, rozhodli sme sa za trendy pokladať položky nakupované vo viac ako *m%* nákupov.

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 20-26.

Témou sa zaoberáme z dôvodu, že nakupovanie používateľov výrazne podlieha sezónnosti, rozličné položky sú v určitých obdobiach kupované viac, inokedy menej. Pre obchodníka je dôležité dokázať obslúžiť všetkých a preto by mal mať možnosť včas doplniť skladové zásoby, prispôsobiť reklamu náladám zákazníkov a podobne. Trendy v nakupovaní je možné predikovať len do určitej miery, maximálne na úrovni typov tovaru (napr. prezutie zimných pneumatík), nie však konkrétnych položiek (prezutie v pneuservise XY). Je však možné identifikovať ich na základe vývoja nákupov a včas tak odhadnúť, ktoré položky sa budú predávať veľa, resp. viac ako v súčasnosti.

V doméne online zľavových portálov sa položky dynamicky menia. Rýchlo vznikajú, zanikajú či predlžujú platnosť. Na tieto zmeny je potrebné dokázať rýchlo zareagovať. Položiek a nákupov je navyše pomerne veľa, čo spolu s predchádzajúcou vlastnosťou vylučuje dávkové spracovanie. Rýchlosť je v tomto prípade kľúčová aby mal obchodník dosť času zareagovať. Preto je potrebné prúdové spracovanie dát.

Prúdové spracovanie je založené na jednoprechodovom prístupe k dátam [7], kedy nie je možné opätovne prechádzať všetky dáta ale nanajvýš určitú najaktuálnejšiu podmnožinu, prípadne agregované štatistiky. Týmto spôsobom dosiahneme výpočtovo efektívne spracovanie, ktoré však vyžaduje upustenie od presnosti výsledku.

V sekcii 2 opisujeme existujúce prístupy identifikácie trendov v prúde dát. V sekcii 3 na základe zistených poznatkov identifikujeme 2 metódy využívajúce posuvné respektíve skáčuce okno. Tieto metódy overujeme a porovnávame v sekcii 4. Zistené poznatky diskutujeme v závere práce, v sekcii 5.

2 Prístupy identifikácie frekventovaných prvkov

Pre identifikáciu frekventovaných prvkov prúde dát, existuje viacero prístupov. Medzi základné radíme počítadlové, skicové a oknové prístupy, ktoré v tejto sekcii stručne opisujeme a diskutujeme z pohľadu vhodnosti pre náš problém.

2.1 Počítadlové prístupy a skicové prístupy

Prvé dve skupiny prístupov sú založené na postupnom prechádzaní dát a pamätaní si počtu nákupov jednotlivých položiek [3]. Oba prístupy a tiež jednotlivé ich algoritmy sa líšia spôsobom pamätania, udržiavania pamäte prípadne zabúdania nedôležitých prvkov, avšak princíp ostáva rovnaký. Trendy sú identifikované za celú históriu, neaktuálne prvky sú evidované s rovnakou dôležitosťou ako aktuálne. Nové trendy sa len ťažko a postupne presadia voči dlhodobým, ktoré sú už známe a teda ich v skutočnosti identifikovať vôbec netreba a podobne. Z tohto dôvodu pokladáme počítadlové a skicové prístupy za nie vhodné pre riešenie nášho problému včasnej identifikácie trendov v správaní sa používateľov online zľavového portálu.

Princípom počítadlových prístupov je vytvorenie určitého počtu počítadiel – dvojíc položka, početnosť (počet počítadiel je zadaný pevne, prípadne je odvodený z počtu existujúcich položiek). Po nákupe položky sa táto pridá medzi počítadlá, prípadne ak sa tam už nachádza, tak počítadlo inkrementuje. V závislosti od použitého

algoritmu sa líši akcia po nájdení novej položky. Pokiaľ sú všetky počítadlá obsadené inými položkami, napr. algoritmus Frequent dekrementuje všetky počítadlá, algoritmus Space-Saving nahradí najmenej frekventovanú položku, pričom zachová jej početnosť [8].

Skicové prístupy zefektívňujú princíp počítadiel tým, že jednotlivé položky ukladajú formou hešovacích tabuliek, čo zvyšuje pamäťovú efektívnosť a znižuje čas prístupu k položkám, čo má zmysel najmä pri spracovávaní prúdov veľkých dát. Medzi najznámejšie skicové algoritmy patria CountSketch, či CountMinSketch [3].

2.2 Oknové prístupy

Pokiaľ pri identifikácii trendov pracujeme len s aktuálnymi dátami, je vhodné použiť tretí prístup, algoritmy založené na oknách. V tomto prípade pracujeme len s obmedzenou podmnožinou najaktuálnejších dát, vďaka čomu je proces spracovania taktiež menej pamäťovo náročný. Existuje viacero algoritmov využívajúcich okná, ktoré sa líšia spôsobom tvorby okien, ich počtom a spôsobom udržiavania.

Najjednoduchším prístupom je skáčuce okno (Landmark Window), ktorý pracuje s dátami v okne od pevne daného časového medzníka. Okno sa postupne napĺňa dátami, v ktorých sa hľadajú trendy. Po dosiahnutí ďalšieho časového medzníka sa okno premaže a dáta sa začínajú zbierať odznova [6]. Výhodou tohto prístupu je, že okno netreba udržiavať a priebežne kontrolovať aktuálnosť dát, stačí len sledovať nastatie časového medzníka. Nevýhodou je, že po premazaní okna sú trendy identifikované na základe malého a nereprezentatívneho množstva dát.

Druhým prístupom je posuvné okno (Sliding Window), kedy sú trendy v ľubovoľnom čase identifikované z dát za určité stanovené obdobie (napr. posledný deň, hodina) alebo počet nakúpených položiek. Výhodou voči predchádzajúcemu prístupu je, že trendy sú vždy identifikované na základe dostatočného časového obdobia alebo počtu dát. Nevýhodou je, že o jednotlivých nákupoch je potrebné si pamätať čas ich vzniku, prípadne poradie, a po posune okna mimo ne ich odstrániť, čo vyžaduje určitú réžiu [5].

Medzi pokročilé oknové prístupy patrí napríklad algoritmus tlmené okno (Damped Window), ktorý je založený na sérii okien zachytávajúcich rozlične vzdialenú minulosť, pričom okná sú uvažované s rozličnou dôležitosťou klesajúcou smerom k starším dátam. Týmto spôsobom je možné simulovať proces zabúdania informácií v čase prípadne poklesu ich dôležitosti [4]. Tento algoritmus zachytáva viac informácií v porovnaní s predchádzajúcimi, avšak za cenu výrazne väčšej pamäťovej a výpočtovej réžie.

3 Návrh metód včasnej identifikácie trendov

Našim cieľom bola včasná identifikácia trendov v prúde dát zachytávajúcich nákupy používateľov online zľavového portálu. V rámci riešenia sme sa zaoberali využitím prístupu založeného na posuvnom okne, resp. skáčucom okne. V oboch prípadoch sme sa zamerali na také riešenie, ktoré bude výpočtovo efektívne (lineárna zložitosť).

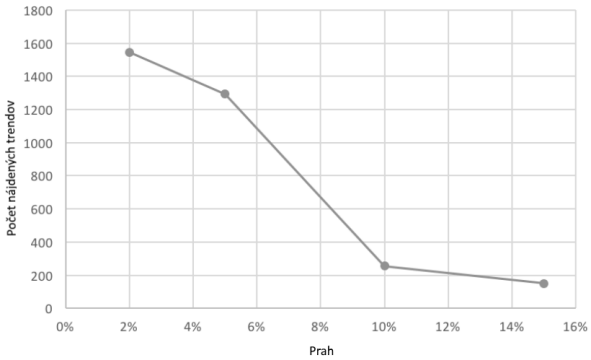
Použité metódy boli v prvej fáze inicializované nastavením vstupných parametrov metódy:

1. Nastavením prahu, pri ktorom nakupovanú položku pokladáme za trend. Presnejšie, sledujeme či podiel nákupov danej položky a celkového počtu nákupov prekonáva stanovený prah.
2. Stanovením veľkosti okna (pre posuvné okno) respektíve vzdialenosti míľnikov (pre skáčuce okno), v rámci ktorých identifikujeme v nákupoch trendy.

Po inicializácii metód sú tieto schopné postupne prijímať prúd dát a spracúvať ho nasledujúcim spôsobom. Pre každý spracovávaný záznam o nákupe:

1. Pokiaľ je časová pečiatka najstaršieho nákupu staršia ako hranica okna (vypočítaná zo spracovávaného záznamu) dekrementuj počítadlo pre túto položku a tú vymaž. Tento krok opakuj pokým nenájdeš najstaršiu položku, ktorú netreba vymazávať.
   a. V prípade skáčuceho okna je postup rovnaký, avšak najstaršia položka sa porovnáva voči poslednému míľniku pred aktuálnou položkou.
2. Inkrementuj počítadlo pre nakúpenú položku.
3. Identifikuj aktuálne trendové položky
   a. Tento krok nie je potrebné vykonávať vždy, ale len na požiadanie, čím sa zníži výpočtová náročnosť algoritmu.

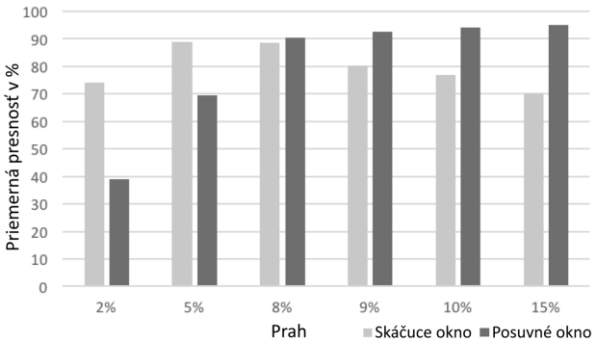
Pri inicializovaní algoritmov sme prah, kedy položku považujeme za trend, nastavili na základe pozorovania, kedy sme zistili, že existuje sigmoidálny pomer medzi hodnotou vzťahu a počtom nájdených trendov (Obr. 1). Na základe tohto zistenia sme experimentovali pri overovaní metódy s viacerými hodnotami zvoleného prahu. Pri vyššom nastavení totiž odfiltrujeme dostatok nezaujímavých položiek, pri nižšom naopak identifikujeme dostatočne širokú základnú položiek, ktoré môžeme neskôr používať pre ďalšie úlohy. Pre účely tohto príspevku sme však zafixovali veľkosť uvažovaného okna na 30 dní, čo zodpovedá typickej dobe ponuky zľavovej položky.



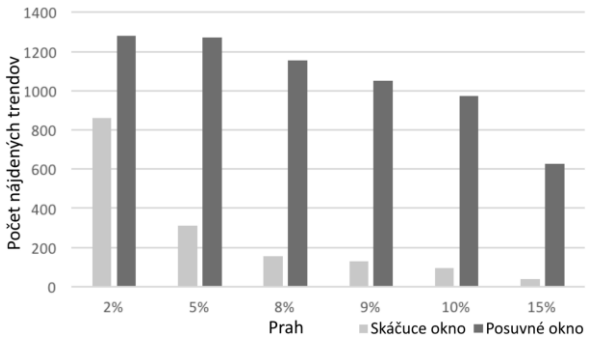
Obr. 1.Závislosť prahu nakupovanosti položky voči ostatným v pomere k počtu položiek (trendov) nakupovaných nad daný prah.

4 Overenie použitých metód

V rámci experimentálneho overenia sme identifikovali trendy v rámci datasetu zo zľavového portálu získaného ako súčasť projektu HIBER [2]. Pracovali sme s 1 mil. záznamov pochádzajúcim z obdobia 6 mesiacov. Tu sme sa zamerali na sledovanie presnosti identifikácie trendov (Obr. 2) a taktiež počet reálne nájdených trendov (Obr. 3). Ako môžeme vidieť, metóda využívajúca posuvné okno dokáže identifikovať výrazne viac trendov ako metóda využívajúca skáčuce okno. Tieto sú však identifikované s menšou presnosťou (okrem prípadov, kedy je prah nastavený na 5% a menej). Metóda skáčuceho okna naopak nedokáže identifikovať toľko trendov, jej presnosť je však vyššia, najme s rastúcim prahom akceptovateľnosti. Dôvodom je to, že táto metóda má k dispozícii menšie množstvo dát a teda na začiatku intervalu vymedzeného medzníkom dokáže definovať len skutočne výrazné trendy.



Obr. 2 Priemerná presnosť identifikácie trendov metódami využívajúcimi okná pri rozlične nastavenom prahu akceptovania trendov.



Obr. 3 Počet nájdených trendov pri rozlične nastavenom prahu akceptovania trendov.

5 Záver

V tomto príspevku sme sa zaoberali problémom včasnej identifikácie trendov v prúde dát. Dostatočne skoro identifikovaná informácia o tom, ktoré prvky budú v krátkej budúcnosti dobre predávané má vysokú hodnotu pre poskytovateľov webových portálov (prípadne akýchkoľvek predajní). Na základe takejto informácie totiž môžu nastaviť svoju predajnú politiky, naskladniť tovar alebo nastaviť výšku zliav.

Na riešenie problému efektívneho spracovania veľkého množstva dát, ktoré ostávajú aktuálne len po obmedzenú dobu, sme použili dva oknové algoritmy. Metóda posuvného okna bola schopná aj na obmedzenej množine dostupných dát identifikovať vysoké percento skutočných trendov. Na druhej strane však za trend označila aj množstvo dát, ktoré trendami neboli. Pri striktnejšom nastavení prahu, pri ktorom boli prvky označované za trend, sme však aj pri tejto metóde dokázali dosiahnuť presnosť 95%. Druhou použitou metódou bolo metóda skáčuceho okna, ktorá aj napriek menšiemu objemu dát, ktoré mala v priemere k dispozícii v porovnaní s predchádzajúcou metódou, dokázala odhaľovať trendy s vysokou presnosťou (70-89%). Počet nájdených trendov bol však pri tejto metóde malý a identifikované boli len veľmi výrazné položky. Jej výhodou je najmä jednoduché udržiavanie dát v okne, keďže stačí sledovať dosiahnutie časového míľnika a následne naraz vymazať všetky dáta. Z hľadiska užitočnosti pre úlohu včasnej identifikácie trendov, však pokladáme za vhodnejšie použiť metódu posuvného okna so striktne zadaným prahom, nakoľko týmto spôsobom dokážeme identifikovať dostatočné množstvo trendov s vysokou presnosťou ich identifikácie.

Literatúra

1. Aggarwal, C. C., Wang, J.: Data Streams: Models and Algorithms. Data Streams, 31, s. 9–38, 2007.
2. Bieliková M. a kol.: Projekt HIBER: hlbšie poznávanie správania sa človeka v digitálnom priestore, WIKT & DaZ 2016, s. 141 – 144, 2016.
3. Cormode, G., Hadjieleftheriou, M.: Finding the frequent items in streams of data, Communications of the ACM, 52(10), s. 97-105, 2009.
4. Gaber, M. M., Zaslavsky, a., Krish-naswamy, S.: Data Stream Mining overview. Data Mining and Knowledge Discovery Handbook, s. 759–787, 2009.
5. Giannella, C., Han, J., Yan, X., Yu, P. S.: Mining Frequent Patterns in Data Streams at Multiple Time Granularities. Next generation data mining, s. 191–212, 2003.
6. Golab, L., DeHaan, D., Demaine, E. D., Lopez-Ortiz, A., Munro, J. I.: Identifying frequent items in sliding windows over online packet streams. Proc. of the 2003 ACM SIGCOMM conference on Internet measurement - IMC '03, s. 173, 2003.
7. Krempl, G. a kol.: Open challenges for data stream mining research. SIGKDD Explor. Newsl. 16(1), s. 1-10, 2014
8. Metwally, A., El Abbadi, A.: Efficient Computation of Frequentand Top-k Elements in Data Streams s. 398–412, 2005.

Poďakovanie: Táto publikácia vznikla vďaka čiastočnej podpore projektov APVV-15-0508 a VG 1/0646/15. Autori článku chcú poďakovať Natálii Čulákovej, ktorej výskum a výsledky bakalárskej práce poslúžili ako základ pre tento článok.

Annotation:

An Early Identification of Trends within Behaviour of Online Discount Portal Users

The paper focuses on an early identification of trends within online discount portal. As the users' behaviour changes in time and they react differently to specific discount, our aim is to say in advance which items will be bought the most. The main challenge is the volume of the data. For this reason we process them as a stream and identify trends by window methods considering only selected subset of the most recent data. We show that in this way it is possible to identify the trends based on first days of their selling.