Analýza prúdu prichádzajúcich udalostí použitím rôznych metód pre analýzu údajov

V súčasnosti pozorujeme vysoký nárast záujmu v oblasti analýzy údajov. Vhodné použitie metód a techník na analýzu prináša hodnotné výstupy pre používateľa, ktoré môžu môžu byť použité pre strategické rozhodnutia v podnikoch. Najčastejší postup je aplikovaním metód ako napríklad lieviková analýza alebo rozhodovacie stromy nad statickou kolekciu dát. Avšak tento prístup neposkytuje okamžitý výsledok pre používateľa. Nedostatkom takýchto prístupov je nutnosť dáta najskôr zozbierať a uložiť, čo je dnes, kedy vznikajú milióny záznamov za deň predstavuje veľký problém.

Pri dolovaní v prúde dát čelíme niekoľkým výzvam: objem, rýchlost (frekvencia) a rozmanitosť. Veľký objem dát, ktoré vznikajú veľmi rýchlo je potrebné spracovať v obmedzenom čase. Pričom veľkosť sa neustále zvyšuje, potenciálne až do nekonečna. Identifikujeme niekoľko najviac zasiahnutých, ktoré sú zdrojmi týchto dát, a to senzory, počítačové siete, sociálne siete a Internet Veci (angl. Internet of Things). Na informácie generované z takýchto zdrojov sa často pozeráme ako na neohraničené prúdy údajov. Spracovanie a následná analýza takýchto prúdov je náročná úloha. Pre aplikácie je kritické spracovať údaje s nízkou odozvou, pričom riešenie musú byť presné a odolné voči chybám. Nakoľko sú prúdy neohraničené vo veľkosti a potenciálne nekonečné, môžeme prejsť len raz cez malý interval prúdu. Potom, dáta musia byť spracované tak ako vznikajú. Tradičné metódy a princípy pre spracovanie statickej kolekcii údajov nie sú postačujúce.

Dolovanie a extrakcia informácií z prúdiacih dát dnes predstavuje niekoľko výziev. Jednou z oblastí je bezpečnosť a dôvernosť v dolovaní dát. Návrh modelu môže byť navrhovaný na pôvodných dátach, ale nasadenie by malo byť realizované na anonymizovaných dátach. Identifikujeme dve hlavné výzvy pre uchovanie bezpečnosti pri dolovaní v prúdoch. Prvou je nekompletnosť informácií, informácie často prúdia nekompletné alebo neaktuálne. Druhá výzva nadväzuje na prvú, a to že dáta sa môžu v čase meniť a vyvýjať. Môže sa meniť štruktúra, komplexita a ich presnosť. Preto, vopred definované bezpečnostné pravidlá a politiky nemusia byť časom aktuálne a pokrývať stanovenú oblasť. Medzi známe techniky dolovania v prúdoch údajov považujeme nasledujúce:

- Klastrovanie, existuje niekoľko výskumov, ktoré sa venovali špeciálne

klastrovaniu implementovaním napríklad k-mediánu a inkrementálnych algoritmov.

- Klasifikácia, opäť existuje niekoľko známych výskumov, ktoré s venujú problému klasifikácie. Napríklad použitím dát z reálneho sveta a umelých dát, pričom implementujú algoritmy, ktoré triedia dáta na základe porovnaní medzi týmito dvoma vzorkami.

- Počítanie frekvencie a opakovaní, použitím posuvných okien a inkrementálnych algoritmov na detekciu vzorov v prúde.

- Analýza časových radov použitím symbolickej reprezentácie šasových radov v prúde dát. Takáto reprezentácia nám umožňuje redukciu veľkosti prenášaných dát. Táto technika pozostáva z dvoch hlavných krokov, aproximácia po častiach a následná transformácia výsledku do diskrétnych veličín.

Predspracovanie prúdu je dôležitým krokom v aplikáciach reálneho sveta. Nakoľko dáta prichádzajú z nehomogénneho sveta, môžu byť zašumené, nekompletné, duplicitné alebo často obsahovať hodnoty, ktoré sa značne líšia od ostatných. Predspracovanie prúdiach údajov je potrebné čo najviac automatizovať. Existuje potreba pre implementovanie metód strojového učenia, ktoré sa adaptujú v čase s meniacimi sa dátami. Avšak, tieto modely a metódy by mali byť synchornizované s prediktívnymi modelmi. Pri aplikovaní strojového učenia a prediktívnych modelov je nutnosť uvažovať historické dáta. Čím podstatne narastá komplexita celého prístupu. Hlavné výskumné problémy rozdeľujeme do dvoch hlavých kategórií, *dátovo orientované* a *orientácia na úlohu*. Dátovo orientované techniky používané pri spracovaní prúdiacich údajov:

- Vzorkovanie je proces výberu dátovej vzorky na spracovanie podľa pravdepodobnostného modelu. Problém pri vzorkovaní je v kontexte analýzy prúdiach dát je potenciálne nekonečný dataset, resp. nevedomosť jeho skutočnej veľkosti.

- Kontrolovanie zaťaženia je proces zahadzovania niektorých, potenciálne nepotrebných vzoriek dát. Opäť tu nástava problém v kontexte prúdu. Môže nastať moment, kedy práve zahodená vzorka dát je pre aktuálnu analýzu príznačná a dôležitá.

- Agregácia je proces vypočtu štatistických údajov nad prúdmi. Problém je pri veľkých tokoch v prúde, pričom je tiež potreba pohľadu do minulosti.

- Aproximačné algoritmy použitie aproximačných algoritmov má za následok podstatné zrýchlenie spracovania a analýzy prúdov za predpokladu istej chybovosti. Chybovosť je zvačsa ohraničená.
- Posuvné okno, tento prístup vznikol s potrebou analýzy definovaného časového okna z prúdiacih údajov.

Medzi hlavné výskumné problémy v tejto oblasti zaraďujeme:

- Nepretržité spracovanie prúdiach údajov.
- Predspracovanie prúdov.
- Aplikovanie správneho analytického modelu a jeho udržateľnosť v čase.
- Formalizácia validácie presnosti spracovania v takmer reálnom čase.
- Dolovanie v prúdoch dát mobilných zariadení a Internetu Vecí.

[1] KREMPL, Georg, et al. Open challenges for data stream mining research. ACM SIGKDD Explorations Newsletter, 2014, 16.1: 1-10.
Diskusia o ôsmich otvorených problémoch pri dolovaní a analýze v prúdoch dát.
[2] GABER, Mohamed Medhat; ZASLAVSKY, Arkady; KRISHNASWAMY, Shonali. Mining data streams: a review. ACM Sigmod Record, 2005, 34.2: 18-26.
Prehľadový článok o teoretických základoch dolovania v prúdoch dát. Známe techniky pre dolovanie a systémy. Aktuálne výskumné vázvy v tejto oblasti.