Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

Matúš Cimerman

Analýza prúdu prichádzajúcich udalostí použitím rôznych metód pre analýzu údajov

Diplomová práca

Vedúci práce: Ing. Jakub Ševcech

máj, 2016

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

Matúš Cimerman

Analýza prúdu prichádzajúcich udalostí použitím rôznych metód pre analýzu údajov

Diplomová práca

Študijný program: Informačné systémy
Študijný odbor: x.x.x Informačné systémy
Miesto vypracovania: Ústav informatiky a softvérového inžinierstva, FIIT STU Bratislava
Vedúci práce: Ing. Jakub Ševcech

máj, 2016

Abstract

Nowadays we can see emerging need for data analysis as data occur. Processing and analysis of data streams is a complex task, first, we particuraly need to provide low latency and fault-tolerant solution.

In our work we focus on proposal a set of tools which will help domain expert in process of data analysis. Domain expert do not need to have detailed knowledge of analytics models. Similar approach is popular when we want analyse static collections, eg. funnel analysis. We study possibilities of usage well known methods for static data analysis in domain data streams analysis. Our goal is to apply method for data analysis in domain of data streams. This approach is focused on simplicity in use of selected method and interpretability of results. It is essential for domain experts to meet these requirements because they will not need to have detailed knowledge from such a domains as machine learning or statistics. We evaluate our solution using software component implementing chosen method.

Abstrakt

Dnes môžeme pozorvať narastujúcu potrebu analyzovať dáta počas ich vzniku. Spracovanie a analýza prúdov dát predstavuje komplexnú úlohu, pričom je dôležité poskytnúť riešenie s nízkou odozvou, ktoré je odolné voči chybám.

V našej práci sa sústreďujeme na návrh súboru nástrojov, ktoré pomôžu doménovému expertovi počas analýzy dát.

Poďakovanie

Na prvom mieste vyslovujem poďakovanie vedúcemu mojej bakalárskej práce, Ing. Jakubovi Ševcechovy, za všetky jeho odborné rady, odovzdané skúsenosti a usmernenie pri tvorení práce.
Touto cestou taktiež vyslovujem poďakovanie všetkým výskumníkom zo skupiny PeWe, za prínosné diskusie a ich spätnú väzbu týkajúcu sa mojej práce.
V neposlednom rade ďakujem celej mojej rodine a priateľom.

Matúš Cimerman

Obsah

1. Úvod

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, veľký problém.

TODO: Pridať vysvetlenie pojmu *analytika*.

Pri dolovaní v prúde dát čelíme niekoľkým výzvam: objem, rýchlosť (frekvencia) a rozmanitosť. Veľký objem dát, ktoré vznikajú veľmi rýchlo je potrebné spracovať v ohraničenom časovom intervale, často v takmer reálnom čase (závisí od kontextu problému). Objem dát sa neustále zvyšuje, potenciálne až do nekonečna. Identifikujeme niekoľko najviac zasiahnutých oblastí, ktoré sú zdrojmi týchto dát, a to senzory, počítačové siete, sociálne siete a Internet Vecí (angl. Internet of Things). Na informácie generované z takýchto zdrojov sa často pozeráme ako na neohraničené a potenciálne nekonečné prúdy údajov. Spracovanie a nasledujúca analýza týchto prúdov je komplexná úloha. Pre aplikácie je kritické spracovať údaje s nízkou odozvou, pričom riešenie musú byť presné, škálovateľné a odolné voči chybám. Nakoľko sú prúdy neohraničené vo veľkosti a potenciálne nekonečné, môžeme spracovať len ohraničený interval prúdu. Potom hovoríme, že dáta musia byť spracované tak ako vznikajú. Tradičné metódy a princípy pre spracovanie statickej kolekcii údajov nie sú postačujúce na takéto úlohy.

2. Analytické úlohy v kontexte prúdu údajov

TODO: Asi bolo vhodné pridať nejaký úvod tejto kapitoly - zamyslieť sa, či to má zmysel.

2.1 Predspracovanie prúdu

Predspracovanie je azda najdôležitejším krokom v aplikáciach reálneho sveta a časovo najnáročnejšou úlohou pre každého analytika. 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 synchronizované s prediktívnymi modelmi. Pri aplikovaní strojového učenia a prediktívnych modelov je nutnosť uvažovať historické dáta. Čím podstatne narastá zložitosť celého prístupu. Hlavné výskumné problémy rozdeľujeme do dvoch hlavých kategórií, *dátovo orientované* a *orientované 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.

2.2 Dolovanie a extrakcia informácií

Dolovanie 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 prichádzajú 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.

2.3 Zhodnotenie

3. Existujúce analytické prístupy a metódy

3.1 Statické metódy

3.2 Adaptívne metódy

4. Otvorené problémy pri analýze prúdu údajov

4.1 Použitie známych metód v prúdoch

4.2 Vizualizácia prúdov a výsledkov analýz nad prúdmi

4.3 Technické výzvy

5. Detekcia trendov v kontexte prúdového spracovania

– navrh mojej metody –
—> tu uz nieco musi byt

- Anomalie
- Eventy
- Trendy
- Sezonnost
- Drilldown nad datami
- PREDIKCIE

6. Zhodnotenie a budúca práca

Literatúra

Príloha A