OPONENTURA BAKALÁŘSKÉ PRÁCE

Autor práce: Tomáš Vyskočil
Téma: Detekce anomálního uživatelského chování

ING. JAN LUKAVSKÝ*

9.6.2015

1 O PRÁCI

V práci se autor zabývá tématem detekce anomálního uživatelského chování z poskytnutých vyhledávacích logů za účelem odhalení chování uživatelů, které by mohlo být označeno za záměrně manipulativní. V prvních dvou sekcích se autor nejprve věnuje motivaci uživatelů manipulovat s prokliky na výsledky vyhledávání, případně na jiné přirozené nebo doplňkové výsledky - například na opravy překlepů, upoutávky, PPC reklamy, a podobně. Popsané motivace uživatelů se dají považovat za poměrně obsáhlé a podchycující valnou většinu reálně prakticky pozorovaných manipulativních chování.

V další kapitole autor shrnuje state of the art metody řešení problému, které se dají běžně dohledat v literatuře, a které byly zároveň následně zvoleny pro implementaci. Z popisu je srozumitelné, jak dané metody fungují, i jakým způsobem je lze aplikovat na autorem řešený problém.

V kapitole (4) popisuje autor výpočetní framework použitý pro implementaci metod z předchozí kapitoly a ačkoliv popis nezabíhá do přílišných detailů, dává i tak čtenáři základní představu o popisovaných systémech.

V následující kapitole jsou hrubě popsána dodaná data, na kterých probíhalo vyhodnocení implementovaných algoritmů, včetně jejich zdroje.

Kapitola (6) se věnuje již konkrétnímu způsobu implementace tří hlavních použitých algoritmů, charakteristikám dodaných data a modifikacím původních algoritmů, které autor vzhledem k dodaným datům zvolil.

Poslední tři kapitoly se věnují jednak popisem implementace použitých algoritmů, následně vyhodnocení a závěru.

2 POUŽITÉ METODY

Autor si pro implementaci vybral dvě metody založené na modelování pravděpodobnosti uživatelské session (sekvence akcí uživatele v SERPu - Search Engine Result Page) pomocí Markových řetězců na základě statistiky podmíněných přechodů mezi různými stavy (zadání dotazu, proklik výsledku,

* *Seznam.cz, a.s., Praha, Česká Republika*

stránkování, apod.) a jednu metodu založenou na modelování uživatelské session v N dimenzionálním prostoru a hodnocení její "nepřirozenosti"definicí vzdálenostní metriky od průměru, tedy od něčeho, co by se dalo považovat za "přirozené"(model dále označovaný jako MFM).

Zatímco první z Markovských metod bere do úvahy pouze přechody mezi autorem definovanými uživatelskými akcemi, které může uživatel vykonal v SERPu, bere druhá metoda do úvahy ještě časový rozměr (diskretizovaný), čímž značně navyšuje počet různých stavů. Rozšíření o časovou složku ve druhé metodě považuji za vhodně zvolené, neboť i přes zvýšení počtu stavů, a tím pádem počtu přechodů mezi nimi, je jejich počet stále únosný a dá se předpokládat, že by neměl být problém získat pro každý přechod dostatečný objem dat, aby nebyla statistika příliš řídká.

Autor provedl poměrně kvalitní vyhodnocení jednotlivých algoritmů na základě primárně automatického hodnocení přes nepřímé ukazatele, které mu byly poskytnuty týmem fulltextového vyhledávače. Metody vyhodnocení považuji vzhledem k datům za vhodné, ačkoliv v souladu s autorem práce bych řekl, že použitá metodika hodnocení může být trochu "hrubá"a její výsledky spíše orientační, než exaktní. I tak je ale z výsledku vidět, že jako samostatný model se nejlépe jeví model MFM. Vzhledem k tomu, že Markovské a MFM modely se jeví být do značné míry nezávislé, považuji za velmi vhodně zvolené řešení "mixture model", tedy kombinace časově závislého Markovského modelu a MFM.

3 HODNOCENÍ PRÁCE

Práce je zpracována velmi kvalitně, veškerá použitá literatura je dobře ozdrojovaná a až na pár velmi drobných problémů s formátováním je výborně čitelná a pochopitelná. Implementace navržených algoritmů ve frameworku MapReduce byla bez pochyb pracná, hlavně pokud jde o správné odladění na netriviálně velkých datech. Zároveň práce poskytuje velmi dobrý materiál pro další studium a navrhuje další možnosti, kudy by šlo závěry dále rozvíjet a zdokonalovat. Z tohoto důvodu hodnotím práci stupněm **A - výborně**.

.....
Ing. Jan Lukavský