ANALYSIS OF DATA ON SOCIAL NETWORKS BASED ON DATA MINING

Marek Fešar

Master Degree Programme (2), FIT BUT

E-mail: xfesar00@stud.fit.vutbr.cz

Supervised by: Vladimír Bartík

E-mail: bartik@fit.vutbr.cz

Abstract: The article explains principles of data mining in the environment of social networks. Focus is put on Twitter microblogging service which provides huge amount of publicly available data suitable for mining. A density based algorithm used for discovering relations among so called hashtags is described.

Keywords: data mining, social networks, density based clustering, tweet, hashtag

1. ÚVOD

Sociální sítě nás obklopují ze všech stran a jen málokterý uživatel Internetu se jim zcela vyhýbá. Aniž bychom si to mnohdy uvědomovali, sdílíme na nich více osobních informací, než na první pohled uvádíme. Cenné informace se často vyskytují skrytě například formou obsahu našich příspěvků sdílených v různých časech na rozličných místech. Až když tyto poskládáme opět dohromady, můžeme dojít většího poznání o konkrétním člověku. Stejně tak se můžeme zabývat komunitou osob, přispívajících ke stejnému tématu. A právě na to se práce zaměřuje.

Disciplína dolování dat se oblasti sociálních sítí věnuje teprve krátce a není mnoho obecně rozšířených algoritmů, které by se uplatňovaly, jako je tomu třeba v oblasti dolování asociačních pravidel (FP strom) či predikce časových řad. Obvykle pracujeme s pojmem sociální graf, což chápeme jako grafovou strukturu, kde množina uzlů (vrcholů) jsou především uživatelé a také další entity sítě – stránky, události, aktivity a jiné (v závislosti na terminologii konkrétní sítě), a hrany pak charakterizují orientované nebo neorientované vazby mezi nimi.

V práci se seznámíme s dolováním na sociálních sítích, zejména pak se sítí Twitter, která nabízí značné množství veřejně dostupných informací vyjadřujících názory svých přispěvatelů. Ukážeme si, jakým způsobem lze v příspěvcích objevovat skryté souvislosti a také jejich následné využití v komerční oblasti.

2. DOLOVÁNÍ V SÍTÍCH

Dolovací algoritmy jsou typicky založeny na grafových algoritmech, neboť právě jim se sociální síť podobá nejvíce. Mezi známé úlohy patří dolování vztahů [1] – klasifikace objektů založená na vztazích, predikce typu objektu nebo vazby, predikce její existence či detekce skupin a podgrafů. Poněkud stranou stojí dolování metadat. Nechceme-li se vydat cestou již poznaného, nezbývá, než se k problematice postavit z vlastního pohledu a s vlastními nápady.

2.1. PROBLEMATIKA SOCIÁLNÍCH SÍTÍ

Sítě nabízejí téměř neomezené množství dat, která každým dnem dále přibývají, aby reflektovaly aktuální postoje svých uživatelů. Dostat se k datům lze obecně dvěma přístupy. Zpracováváním zdrojového kódu stránek vrácených serverem, podobně jako ho zpracovává webových vyhledávač,

nebo prostřednictvím dotazů na dostupné programové rozhraní sítě (API – Application Programming Interface).

Zdrojový kód má povahu obecnější šablony, ve které se v sémanticky rozdílných značkách vyskytují útržky informací, jež musíme poskládat do správného kontextu zpracováním celé stránky. Pozornost vždy odvádějí nevýznamové prvky stránky jako reklamní sdělení, navigace a podobně. Výhodou nám může být nástroj schopný s menšími úpravami pracovat nad stránkami různých sítí.

Naproti tomu dotazování programového rozhraní je více specifické pro potřeby jedné sítě, avšak vrácená informace je daleko přesnější a obsahuje zpravidla jen poptávané informace. V práci s rozhraním sítí jako Facebook, Twitter nebo LinkedIn jsou určité rozdíly z hlediska technického i logického přístupu. Technicky se liší zpravidla formátem dotazů a odpovědí, kdy dotazy přijímá většina prostřednictvím HTTP protokolu (REST API) a odpovědi bývají strukturovány jako XML dokumenty či JSON objekty. Logická diference je daná tím, co síť obsahuje a jakým směrem má smysl dotazy vést – zda chceme objevit kariérní historii osoby na profesně orientovaném LinkedIn, nebo nás zajímají vyjádření na mikrologu Twitter.

Častým omezením v zisku dat pro dolování je limit v počtu dotazů na programové rozhraní za jednotku času a také omezení přístupu k informacím pomocí systému přístupových práv udělovaných jednotlivými uživateli k jejich osobním informacím. Zde je čestnou výjimkou právě síť Twitter, kdy většina příspěvků zvaných *tweety* je volně přístupná.

2.2. DOTAZOVÁNÍ PROGRAMOVÉHO ROZHRANÍ TWITTERU

Twitter poskytuje prostřednictvím svého API přístup prakticky ke všemu, co lze číst i webovým prohlížečem (což není u ostatních pravidlo). Pokud se vyrovnáme s omezením na počet dotazů za čas (obvykle 15 minutové časové okno), obdržíme z něj takřka libovolné tweety, které často glosují aktuální společenskou problematiku, tedy to, co uživatele zajímá. A co zajímá uživatele, by mělo zajímat i nás, zejména s ohledem na obvyklé marketingové využití vydolovaných dat.

V oblasti dotazování směřuje naše pozornost ke speciálním značkám užívaným v jednotlivých tweetech zvaným *hashtagy*. Jedná se o kategorizující slovo, které není dáno žádným pevným omezujícím výčtem, nýbrž může být uživatelem libovolně vybráno. Takový slov pak může jeden tweet obsahovat neomezeně (v rámci 140 znaků). Na základě hashtagů je umožněno vyhledávat další příspěvky uživatelů, neboť existuje předpoklad, že hashtag vystihuje problematiku, jíž se tweet věnuje.

Právě vícenásobné použití hashtagů napříč sítí a zároveň rozdílné hashtagy v rámci jednoho tweetu vytvářejí určitou tematickou síť, jíž se budeme zabývat. Vycházíme z předpokladu, že uživatel *tweetuje* k nějakému tématu a zároveň čte další příspěvky k onomu tématu nabídnuté sítí na základě stejných hashtagů. Užíváním více značek pak uživatel nevědomky sděluje, jaká témata ho oslovují a mohou spolu souviset – z pohledu nejen jeho, ale i dalších uživatelů. A pokud si uživatel čte příspěvky obsahující jím používané značky, můžeme si je skrze rozhraní číst i my.

2.3. ALGORITMUS SBĚRU DAT

Algoritmus pro sběr začíná od prvního (iniciálního) hashtagu zadaného uživatelem při vytváření dolovací úlohy. Ten si přidá do seznamu dosud nevyšetřených hashtagů pro úroveň 1. Iterativně se odebírají hashtagy ze seznamu v dané úrovni následovně. Dokud je seznam nevyšetřených hashtagů aktuální úrovně neprázdný, vezme se první položka, přesune se do společného seznamu vyšetřených hashtagů a provede se dotaz na programové rozhraní Twitteru, který vyhledá všechny tweety obsahující daný hashtag. Vrácené tweety jsou postupně vyšetřeny na přítomnost hashtagů a ty jsou vždy přidávány do seznamu dosud nevyšetřených hastagů pro úroveň $n+1$ tehdy, pokud se nenacházejí ve společném seznamu již vyšetřených hashtagů. Tímto se zabraňuje smyčkám v dotazování. Pro každý obdržený tweet se navíc vytvoří záznam v databázi, čímž se připravují data pro následné dolování. Po vyprázdnění seznamu nevyšetřených hashtagů se inkrementuje počíta-

dlo úrovně. Algoritmus skončí, jakmile narazí na maximální zadanou úroveň. Pro 2 úrovně lze takto nasbírat 8000-10000 hashtagů při potřebném počtu 160-220 dotazů.

2.4. POPIS NAVRŽENÉHO DOLOVACÍHO ALGORITMU

Objevování shluku v nasbíraných datech začíná od iniciálního hashtagu. Algoritmus vychází z metody DBSCAN [2], která je původně zamýšlena pro odhalování shluků v prostorových datech na základě hustoty objektů. V nasbíraných datech se zjišťuje, v jakých tweetech se hashtag vyskytoval a s kterými jinými hashtagy byl užit v každém z nich. Takto vznikají dvojice, z nichž pro každou si algoritmus uchovává násobnost vztahu – kolikrát se hashtagy vyskytly v různých tweetech spolu. Algoritmus uvažuje jen vztahy s násobností, která překračuje zvolenou hranici ε. Pokud má hashtag alespoň *MinPts* vztahů s jinými hashtagy, označíme jej jako jádro. Sousedy jádra pak dále vyšetřujeme na to, zda se jedná taktéž o další jádra, či nikoli. Ve chvíli, kdy nelze nalézt další jádro, rozšiřování shluku končí.

V terminologii DBSCAN objevujeme množinu objektů (hashtagů), které jsou spojeny na základě hustoty. Shluk se tedy rozšiřuje tak dlouho, dokud hustota neklesne pod mez danou parametry ε a *MinPts*. Aby nebylo nutné specifikovat přesné hodnoty obou parametrů, což je považováno za nevýhodu jakékoli dolovací metody, inspiroval se algoritmus optimalizací OPTICS [3]. Ta dovoluje odhalit z předpočítaných hodnot parametrů ε a *MinPts* všechny shluky takové, jejichž *MinPts'* je větší nebo rovno *MinPts* bez nutnosti prohledávat celý soubor dat znovu.

Dle naměřených výsledků zabere dolování v experimentálním souboru přibližně 9000 hashtagů jednotky sekund. S uvedenou optimalizací se pak rychlost posouvá o dva řády, tedy na úroveň setin. Zrychlení lze považovat za rozhodující z hlediska vstřícnosti k uživateli.

3. UKÁZKA ZÍSKANÉ ZNALOSTI

Hashtagy nalézající se v jednom shluku s iniciálním tagem lze z hlediska zájmu uživatelů považovat svého druhu za synonyma. Čím silnější vazba mezi nimi je, tím větší množství uživatelů oslovují. V oblasti marketingu pak objevené tagy využijeme v komerčních příspěvcích, jejichž cílem je oslovit potenciální zákazníky, nicméně ne příliš *průhledným* způsobem. Na zadaný tag „pivo“ tak algoritmus vrací zdánlivě nesouvisející „farmville“ a komerční sdělení propagující výrobek lze na základě získané znalosti formulovat: „Naše #pivo Vám zpříjemní chvilky s #farmville“, což může přivést nové návštěvníky na Twitter účet našeho pivovaru.

4. ZÁVĚR

Práce ukazuje možnosti dolování na sociální síti Twitter za použití programového rozhraní poskytovaného sítí. Je popsán směr, jakým vést dotazy k zisku potenciálně užitečných dat použitelných pro vlastní navržený dolovací algoritmus. Inspirací se stalo dolování založené na hustotě, určené primárně pro prostorová data. Cílem algoritmu je poskytnout základ pro objevování informací použitelných v oblasti marketingu.

REFERENCE

[1] Han, J., Kamber, M.: Data mining: Concepts and techniques, San Francisco, Morgan Kaufmann Publishers, second edition, 2006, ISBN 80-85615-77-0, 770 s.

[2] Zendulka, J., Bartík, V.: Získávání znalostí z databází: Studijní opora, Brno, Vysoké učení technické, 2009, 160s.

[3] Ankerst, M., Breunig, M., Kriegel, H.-P.: OPTICS: ordering points to identify the clustering structure, ACM SIGMOD Record, 1999, s. 49-60