DATAMINING OF RELEVENAT INFORMATION FROM WEB WITH USING SOCIAL NETWORKS

Jakub Smolík
Master Degree Programme (2), FIT BUT
E-mail: xsmoli06@stud.fit.vutbr.cz

Supervised by: Jan Samek
E-mail: samejan@fit.vutbr.cz

Abstract: We did some research regarding web mining and its aspects. There we highlighted major points about importance of web mining, its main purpouses and usage. We also disclosed three main file formats in whose are internet data mainly stored - HTML, XML, RSS. Upon these informations we began to suggest search engine kind of application, with ability to obtain relevant data from the internet using multiple data sources. One of the main sources are going to be social network RSS feeds. Then we made some points on how to extend its utilization in area of trust modeling.

Keywords: data mining, web mining, relevant informations, social networks, information sources

1 ÚVOD

Pod pojmem dolování dat se skrývá věda objevování nových informací. Může se jednat o neznámé vzorce nebo skryté vztahy, které se mohou v obrovském množství dat vyskytovat, i když to nemusí být na první pohled patrné. Jde nám o získávání dříve neodhalených znalostí. Cíl naší práce tak spočívá ve vytvoření aplikace, která bude využívat moderní technologie k automatizaci získávání a analýzy vybraných dat, jež jsou generována dnes a denně milióny uživatelů internetu, především pak v prostředí sociálních sítí. Na základě analýzy těchto údajů pak lze najít relevantní informace, určit různé aspekty a trendy v sociálních sítích, důvěryhodnost nebo relevanci jednotlivých zdrojů dat, provádět sociální a statistické výzkumy, apod.

2 VÝZNAM A POUŽITÍ WEB MININGU

Abychom pokryli informační zázemí pro vývoj naší aplikace, prozkoumali jsme problematiku Web miningu, což v překladu znamená dolování informací z webů. V praxi se však spíše používá anglická varianta tohoto termínu, proto se jí v následujícím textu přidržíme. Web mining posouvá prostředí WWW směrem k použitelnější a dostupnější variantě, kdy mohou uživatelé rychle a snadno nacházet potřebné informace. Pomáhá tak aktivně k nárůstu efektivity při jejich získávání, protože zvyšuje relevantnost zkoumaných dat, stejně jako snižuje dobu potřebnou pro jejich nalezení. Hlavním cílem je tedy získat data, která můžeme posléze analyzovat, například pomocí technik data miningu. Data mining je obecné označení analytické metodologie získávání netriviálních skrytých, nebo potenciálně užitečných informací z dat [1].

Nyní si oblasti využitelných informací rozdělíme a uvedeme příklady pro každou z nich:

2.1 WEB CONTENT MINING

Jedná se o hlavní třídu z oblasti web miningu, kdy se soustředíme na vlastní sběr dat z obsahu stránek. V zásadě jde o extrakci textu, obrázků, grafů, tabulek a dalších částí webu, z kterých se poté určí

relevance získaných dat na základě vyhledávacích kritérií. Právě tato oblast bude tvořit hlavní náplň funkcionality námi tvořené aplikace.

Pro zvýšení relevance vyhledávaných informací poskytneme uživateli možnost hodnocení výsledků hledání v kontextu informačních zdrojů, což povede k možnosti personalizace vyhledávacího algoritmu.

2.2 USAGE MINING

Pod tímto názvem se skrývá dolování přidružených statistik ohledně webových stránek, většinou tedy logů. Lze zkoumat návštěvnost, pohyb uživatelů na sledovaných portálech, příchozí a odchozí adresy, doby pobytu na stránkách, apod. Dále nám poskytuje prostředky pro posouzení efektivity reklamy, nebo použitelnosti webového prostředí.

2.3 STRUCTURE MINING

Prostředí formátu HTML postrádá standardizaci, co se struktury obsahu týče, a tak je nutné si poradit jinak. Proto vznikla třída dolování struktury, kde se snažíme shlukovat webový obsah se stejnou strukturou, na který pak lze aplikovat podobná extrakční pravidla pro získání obsahu z nich. Vznikají nám tak různé skupiny stránek, obvykle od stejného autora, nebo ze stejného portálu, mezi kterými lze obsah snadno porovnávat. To může usnadnit určování relevantnosti zkoumaných dat.

2.4 INTERNETOVÝ MARKETING

V hlavní řadě je však jakékoliv dolování informací z internetu především nástrojem pro marketing. Poskytuje totiž velké množství informací o potencionálních zákaznících, konkurenčních výrobcích, vývoji na trhu, odhadu směřování trendů, atd. Využití se nachází především u reklamy, komunikace se zákazníkem, získávání zpětné vazby, šíření povědomí o společnostech a spoustě dalších oblastí. Tyto fakta lze aplikovat při případné komercializaci naší aplikace.

3 ZKOUMANÉ FORMÁTY DAT - HTML, XML, RSS

Web mining je primárně zaměřen na dolování informací z internetového prostředí, kde je nejpoužívanějším formátem souboru HTML. To z něj činí pravděpodobně největší zdroj informací na světě. Problémem však zůstává heterogenita, nekonzistence a různá úroveň strukturovanosti webů, což jejich automatické zpracovávání značně znesnadňuje, což dosvědčuje velký vědecký zájem v dané oblasti [2]. Proto zpracování dat z HTML prozatím z návrhu aplikace vynecháme.

XML se uchytil jako jeden z nejvyužívanějších formátů pro uchovávání a přenos informací na internetu. Pro analýzu byl námi zvolen především díky strukturovanosti formátu, kdy lze zautomatizovat získávání dat z něj [3]. Budeme s ním počítat jako s potencionálním zdrojem informací při tvorbě aplikace.

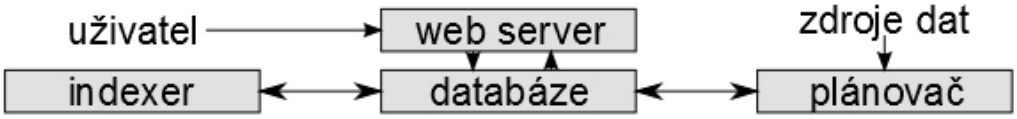
RSS je představitelem z rodiny formátu XML, který slouží pro prezentaci informací, jako jsou například nejnovější články, aktualizace, výsledky hledání, a jejich sdílení s ostatními uživateli [4]. Na rozdíl od obecného standardu XML se RSS vyznačují v rámci jednotlivých verzí jednotnou strukturou. Navíc dovoluje získat data ze sociálních sítí, což skrývá potenciál, který hodláme při vývoji využít.

4 ZDROJE A APLIKACE

Jako zdroje dat hodláme použít jak fixní dostupné kanály RSS z internetových portálů, tak sociální sítě, z kterých lze data automaticky extrahovat. Ačkoliv většina sociálních sítí úmyslně neposkytuje

volně dostupné informace týkající se její struktury, lze v nich užitečné datové kanály v mnohých případech odhalit [2]. Pro nás zajímavou se stala možnost tvorby RSS kanálů z výsledků hledání klíčových slov, nebo podle tagů u příspěvků a dokumentů. S využitím full-textového vyhledávání v přirozeném jazyce, které výsledkům přiřazuje míru relevance, budeme vyvářet index klíčových slov ze zvolený zdrojů.

Pro implementaci bude použit jazyk PHP ve spojení s databázovým serverem MySQL. Aplikace bude rozdělena na dvě části, z nichž jedna bude poskytovat komunikaci s klientem, zatímco druhá bude samostatně provádět průběžnou aktualizaci databáze na základě informačních kanálů od vybraných zdrojů a nalezené shody ukládat, jak je uvedeno ve schématu na Obrázku 1.



Obrázek 1: Schéma aplikace.

Uživatel po přihlášení do systémů zadá úkol najít hledaný výraz ve vybraných informačních zdrojích, mezi které bude moci přidávat i vlastní definované zdroje. Vyhledávání klíčových slov bude probíhat s využitím optimalizovaných regulárních výrazů s upravitelným nastavením. Jako výsledky se mu poté zobrazí články obsahující shodu, informace o nich a odkazy na zdroje. Samotný proces hledání tak bude dlouhodobější a bude probíhat od momentu spuštění hledání do budoucnosti, což zaručí aktuálnost informací.

Na základě analýzy výsledků vyhledávání a na základě relevenace nalezených klíčových slov bude system umožňovat stanovení vhodnosti sledovaného zdroje pro zvolena klíčová slova. S využitím zpětné vazby od uživatelů systému, kdy každý uživatel může hodnotit výsledky hledání, bude možné navíc personalizovat vyhledávací algoritmus a upřednostňovat výsledky z preferovaných zdrojů."

5 ZÁVĚR

Provedli jsme průzkum v oblasti zabývající se data a web miningem, z které jsme shrnuli nejdůležitější poznatky z daných oblastí. Na jeho základě jsme navrhli aplikaci, jež se zaměří na získávání relevantních dat z různých informačních zdrojů dostupných na internetu. Cílem bude dodat uživateli výsledek, který poskytne komplexní pohled na vyhledávaný objekt, přičemž následná analýza výsledků hledání, může pomoci stanovit relevantnost a hodnocení jednotlivých informačních zdrojů. Tyto data lze v budoucnu využít pro výzkum zabývající se tvorbou sítí důvěry na internetu.

PODĚKOVÁNÍ

Tato práce byla částečně podpořena operačním programem Výzkum a vývoj pro inovace v rámci projektu Centrum Excelence IT4Innovations (CZ.1.05/1.1.00/02.0070).

REFERENCE

[1] Scime, A.:Web mining. Idea Group Publishing, 2005, iSBN 15-914-0414-2.

[2] Bing, L.:Web Data Mining. Springer, 2011, iSBN 978-3-642-19459-7.

[3] Nayak; Richi:XML Data Mining: Process and Applications. Idea Group Inc. / IGI Global, 2008.

[4] Cover, R.: RDF Rich Site Summary (RSS).http://xml.coverpages.org/rss.html, 2007.