Acta Informatica Pragensia
3(3), 2014, 280–287, DOI: 10.18267/j.aip.54
Online: aip.vse.cz

Sekce / Section:
Recenzované stati / Peer-reviewed papers

Nástroje pro automatické rozpoznávání entit a jejich vztahů v nestrukturovaných textech

Tools for Automatic Recognition of Persons and their Relationships in Unstructured Data

Jaroslav Ráček[1], Jan Ministr[2]

[1] Techniserv, spol. s r.o.
Moskevská 949/86, 101 00 Praha 10

[2] Katedra aplikované informatiky, Ekonomická fakulta,
Vysoká škola báňská – Technická univerzita Ostrava
Sokolská třída 33, 701 21 Ostrava 1

jracek@techniserv.cz, jan.ministr@vsb.cz

Abstrakt: Článek se zabývá specifiky třívrstvé architektury softwarového systému, který využívá automatické nástroje pro rozpoznávání a identifikaci osob, objektů a vztahů v nestrukturovaných datech. Datovou vrstvu tvoří modul pro pořizování dat a modul pro správu a vyhledávání v datech. Aplikační vrstva je tvořena samostatnými moduly, které lze kombinovat pro potřeby konkrétních vyšetřovacích policejních úloh. Prezentační vrstva zpřístupňuje výsledky analýz. Celkové řešení demonstrováno na vyvíjeném systému ARIO, které je aplikovatelné pro řadu státních i mezinárodních institucí, ale i soukromých subjektů.

Klíčová slova: Nestrukturovaná data, Monitorování internetu, Identifikace objektů a vztahů, Policejní informační systém, Plugin.

Abstract: The article deals with the specifics of the three layer architecture of software for automatic detection and identification of persons, objects and relationships in unstructured data. The data layer consists of data acquisition and data management modules. The application layer is composed of separate modules that can be combined to meet the needs of specific investigative tasks. The presentation layer makes the analysis of results for police investigation. The overall solution is demonstrated on develop the system ARIO which is applicable for a range of national and international institutions as well as private entities.

Keywords: Unstructured Data, Internet Monitoring, Identification of Objects and Relationships, Police Information System, Plugin.

1 Úvod

Tento článek je věnován softwarovým nástrojům pro sledování, identifikaci, stahování a kategorizaci obsahu různých internetových zdrojů, na jejichž vývoji autoři pracují v rámci řešení výzkumného projektu ARIO - Automatické rozpoznávání a identifikace objektů v internetu. Jedná se o rozsáhlou sadu knihoven funkcí a dalších softwarových komponent, jejímž účelem je identifikovat v internetovém a intranetovém obsahu různé entity, jejich vzájemné vztahy a jejich vliv na okolní svět, jak konstatují Aggarwal a Subbian (2014).

Konkrétní nástroje, o kterých hovoří tento článek, začaly původně vznikat pro potřeby komerčních firem, např. pro marketingové účely v sociálních sítích a na diskusních fórech. Tam se využívaly pro identifikaci diskuzí vztažených k jednotlivým produktům, hodnotily sentiment (spokojenost) jednotlivých diskutujících a identifikovaly konkurenční produkty. Ukázalo se však, že vyvíjená funkcionalita je natolik silná, že ji lze s úspěchem využít i pro potřeby bezpečnostních složek při pátrání po informacích o předmětech, osobách a událostech, které se v internetovém obsahu vyskytují. Proto byl zahájen další vývoj a rozvoj funkcionality existujícího softwaru. Řada nástrojů byla rozšířena a současně vznikaly i nové moduly určené primárně pro využití policií, jejíž dekomponovanou funkcionalitu blíže rozebírají Xu a Chen (2005).

V praxi to znamená, že například při pátrání po odcizených uměleckých dílech, jsou tyto nástroje schopny monitorovat strukturovaná nebo částečně strukturovaná data, jako jsou nabídky aukčních síní, elektronické obchody a inzerce, ale i nestrukturovaná data jako jsou diskusní fóra a sociální sítě. Získané informace je třeba následně ukládat do nově vytvořeného pracovního prostoru, nad kterým pracují další analytické a vyhledávací nástroje.

Ačkoli se jedná převážně o práci s textovými daty, součástí celkového řešení jsou i nástroje pro identifikaci objektů v obrazech a videu, přičemž software rozpoznává celé objekty i jejich fragmenty. Části pracující s textovými daty jsou schopny pracovat s texty v různých evropských jazycích.

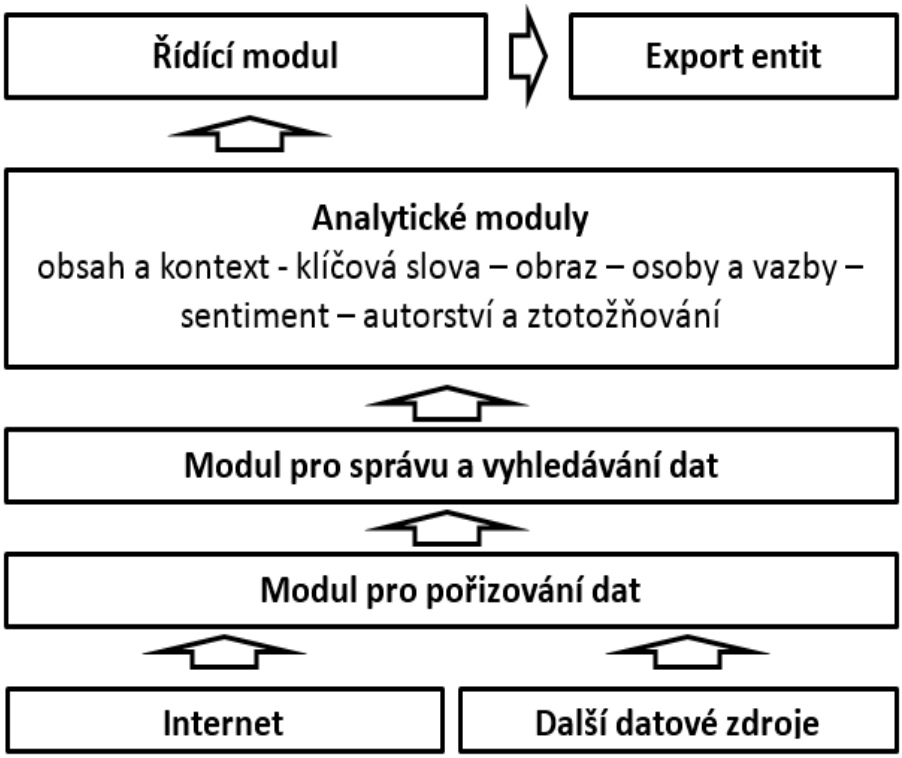
2 Koncepce řešení architektury systému

Celkové softwarové řešení systému, který využívá automatické nástroje pro rozpoznávání a identifikaci osob, objektů a vztahů v nestrukturovaných datech se skládá ze čtyř hlavních skupin nástrojů, jimiž jsou:

- nástroje pro sledování a stahování obsahu internetu,
- nástroje pro správu a vyhledávání uložených dat,
- nástroje pro analýzu dat,
- nástroje pro prezentaci výsledků.

Celý systém funguje tak, že jsou automaticky sledovány vybrané internetové adresy a v případě, že se narazí na zajímavá data, jsou tato data uložena na lokální server. K datům staženým z internetu se dále přidávají data z dalších datových zdrojů, mezi které patří například emailové komunikace a dokumenty pocházející z počítačů a účtů zájmových osob. Data z této vrstvy jsou následně analyzována a z výsledků jsou sestavovány reporty pro koncové uživatele. Celková architektura takto pojatého modulárního řešení systému je ukázána na obrázku č. 1. Obrázek ukazuje základní vrstvy a směr zpracování dat mezi moduly na konkrétním řešení systému ARIO (Automatické rozpoznávání a Identifikace objektů). Nejnižší datovou vrstvu představuje internet a další datové zdroje, což jsou např. emaily a

dokumenty z jednotlivých počítačů nebo data z pasivních sond. Tato data importuje do systému modul pro pořizování dat. Následně jsou data ukládána v jednotném formátu pomocí modulu pro správu a vyhledávání dat. Vedle ukládání tento modul zajišťuje i první analytickou podvrstvu druhé aplikační vrstvy, kterou je vyhledávání zájmových dat s cílem zúžit jejich množství tak, aby nad nimi mohly následně pracovat specializované analytické moduly, jejichž základní funkcionalitu popsal a kategorizoval Sathi (2012). Typicky se jedná o výběr dat podle zdroje, času vzniku, času pořízení nebo autora.



Obr. 1. Architektura systému ARIO.

Takto zúžená data jsou pak podrobována hlubším analýzám, které zajišťují specializované analytické moduly aplikační vrstvy. Parametry analýz a jejich kombinace jsou řízeny řídícím modulem, který výsledky zobrazuje, ale i předává v podobě striktně definovaných entit do dalších systémů policie.

3 Proces pořizování a správy dat

Z pohledu třívrstvé architektury lze chápat jako datovou vrstvu systému ARIO dvojici modulů pro pořizování dat a pro správu dat. Funkcionalita takto pojaté datové vrstvy vychází z popisu metod, které blíže definuje Ashton et al. (2014). Procesy, které probíhají na datovou vrstvou, lze pak rozdělit do dvou následujících skupin.

3.1 Proces sledování a stahování obsahu internetu

Pro pořizování dat z internetu neexistuje jedna univerzální technika. Postup získávání dat je třeba volit v závislosti na typu stránek, ze kterých se data pořizují. Zjednodušeně lze říci, že jiné techniky se používají v případě sociálních sítí a jiné u zbytku internetu, jako jsou inzerce a diskuze. V případě diskuzí, inzercí a dalších jim podobných stránek se postupuje

tak, že vybrané části, typicky jednotlivá diskusní fóra, se stahují celá, respektive se stahují v dostatečně krátké časové periodě. Všechny nové příspěvky se ukládají na vyhrazený server. K tomu lze použít technologii Web Harvest nebo Heritrix, která má univerzálnější použití. Pro většinu serverů je však plně dostatečná technologie Web Harvest.

V případě sociálních sítí je situace odlišná. Techniky Web Harvest nebo Heritrix zde příliš nefungují. Zde je vesměs třeba obsah serveru získávat pomocí speciálních pluginů, které automaticky procházejí ať již veřejnou nebo neveřejnou část sítě, do které se zpravidla dá proniknout pomocí k tomu speciálně zřízených uživatelských profilů. Úspěšnou technikou je také u některých typů sítí online sledování a ukládání aktualit a krátkých zpráv uživatelů, jak uvádí Khan et al. (2014).

Vedle těchto postupů lze přidávat do databáze i další data, která nejsou získávána výše uvedenými technikami. Jako příklad takových dat uvádíme import mailů nebo data z pasivních sond pro sledování provozu telekomunikačních sítí. Kombinací výše uvedených zdrojů vznikají rozsáhlé datové soubory, z nichž lze při vhodné analýze získat velmi cenná data.

3.2 Proces pro správu a vyhledávání uložených dat

Vzhledem k tomu, že získaná data jsou rozsáhlá, není vhodné je ukládat v klasických relačních databázích. V závislosti na povaze a množství pořizovaných dat, ale také na množství prostředků, které jsou k dispozici, se využívají k těmto účelům specializované servery.

Systém ARIO pro tyto účely používá kombinaci nonSQL databáze MongoDB, relační databáze PostgreSQL a indexačního nástroje Apache Solr. Výhodou použití těchto nástrojů je jejich kapacita, rychlost a schopnost indexace nestrukturovaných dat. Zároveň se pomocí těchto nástrojů řeší i prvotní filtrace dat ještě před tím, než je provedena vlastní analýza. Z pohledu analýzy to znamená, že na data jsou zpravidla nejdříve aplikovány funkce, které má v sobě zabudované Apache Solr, čímž je proveden první stupeň analýzy a následně jsou na takto zúženou množinu dat aplikovány specializované analytické algoritmy.

4 Funkcionalita aplikační vrstvy zaměřená na analýzu dat

Vytvořené analytické funkce zpravidla pracují s již částečně přetříděnými daty, což v praxi znamená, že data jsou nejčastěji přetříděna podle času, zdroje a typu. Na úrovni práce s databází v těchto systémech je možné aplikovat více analytických funkcí, jako je filtrace podle vybraných slov nebo autora, nicméně tyto možnosti se zpravidla nevyužívají a ponechávají se až na pozdější fáze analýzy, které se pak provádí pomocí dodatečně vyvinutých nástrojů, které vycházejí z potřeb konkrétní oblasti využití systému. Hlavním důvodem je, že tyto filtrace dat je třeba kalibrovat v závislosti na příslušné doménové oblasti, což komerční produkty neumožňují.

Základními úlohami při analýze vybraných a částečně předzpracovaných dat je:

- rozpoznat hledanou entitu, tj. předmět, osobu nebo událost,
- rozpoznat vztahy mezi nalezenými entitami,
- rozpoznat vztah nalezené entity nebo skupiny entit k okolnímu světu.

Tyto úkoly se řeší pomocí speciálních analytických nástrojů, které je třeba vytvořit. U systému ARIO se jedná o několik knihoven funkcí, které se vzájemně kombinují a s jejich pomocí jsou řešeny základní analytické případy, jejichž specifikace vychází ze základních

uživatelských požadavků na daný systém. Charakter základních analytických úlohy je popsán v následujících odstavcích.

4.1 Analýza obsahu a identifikace entit

Analýza obsahu je základní úlohou. Cílem je identifikovat klíčové výrazy, o kterých se v textu mluví. Při této analýze je třeba vyřešit několik základních problémů. Je třeba rozpoznat jazyk, ve kterém je text napsaný, aby se následně jednotlivé části textu již mohly zpracovávat za využití slovníků a pravidel příslušného jazyka. V této části je také třeba vypořádat se s případnými překlepy a nespisovnými výrazy. Následně jsou z textu separována slova nesoucí významovou informaci. Zjednodušeně lze říci, že se jedná o kombinace vybraných podstatných a přídavných jmen, sloves a číslovek. Spolu s tím probíhá analýza synonym, na jejímž základě jsou pak výrazy stejného významu nahrazeny jedním vybraným reprezentantem. Odtud je pak odvozeno téma jednotlivých textů, což může mít například podobu seznamu klíčových slov.

V případě, že se v textu hledá nějaká entita na základě referenčního vzorku, je referenční vzorek porovnáván s identifikovaným obsahem. Výsledek má pak podobu odkazu na místo, kde se hledaná entita vyskytuje a číselné (relevantní) vyjádření míry shody hledaného a nalezeného vzorku dat.

4.2 Analýza obrazových dat

Pro analýzu obrazu se používají externí knihovny třetích stran. Oproti klasickým technikám rozpoznávání obrazu však jako vstup pro tyto analýzy neslouží pouze hledaný obraz, ale vstupuje tam i sémantická textová informace, která popisuje, co je v obraze hledáno a co je na prohledávaném obraze. Algoritmy rozpoznávaní obrazu pak díky této informaci mohou scénu lépe rozdělit a hledaný předmět rozpoznat s větší přesností, než bez příslušné sémantické informace na vstupu. Vstupní sémantické informace mají různou strukturu v závislosti na prohledávané předmětné doméně a zdroji dat. Typickým případem užití je pátrání po odcizených uměleckých dílech, kdy je k dispozici jednak slovní popis, ale i fotografie hledaného předmětu.

4.3 Analýza sociálních vazeb

V případě, kdy jsou v datech rozpoznány osoby, ať již jako autoři nebo osoby, o kterých se mluví, je jednou z klíčových úloh rozpoznat vztahy mezi nimi. Ze záznamů o osobách z různých zdrojů, je tak rekonstruována interní „sociální“ síť osob. U každé osoby jsou evidována témata, v souvislosti se kterými je zmiňována, a osoby se kterými je ve vztahu. U vztahů mezi osobami jsou dále rozlišovány typy vztahů a jejich intenzita. U všech těchto údajů je jako samostatný parametr evidován čas. Díky tomu lze v čase sledovat, jak se vyvíjel předmět zájmů konkrétní osoby, nebo jak se vyvíjely vzájemné vztahy mezi skupinou osob, jak podrobně popisuje Scott (2000). Na základě toho lze osoby segmentovat do skupin nebo stanovit metriky určující blízkost jednotlivých osob. Tato funkcionalita může být například součástí nástrojů pro odhalování organizovaného zločinu.

4.4 Analýza autorství a ztotožňování

Osoby vyskytující se na internetu používají celou řadu pseudonymů nebo vystupují často zcela anonymně. Koppel et al. (2012) konstatují, že je velmi důležité pak je podle rozpoznat, že za více identitami se skrývá stejná osoba. To lze částečně rozpoznat ze strukturovaných

dat, jako je shodné telefonní číslo nebo email, nicméně ve většině případů tyto údaje nejsou k dispozici. K těmto účelům lze však použít údaje z monitoringu výskytu jednotlivých osob, kdy je sledována intenzita, doba a frekvence aktivit jednotlivých pseudonymů. V případě, že několik pseudonymů vykazuje velmi podobné rysy chování, jsou následně spuštěny další analytické funkce, které mají za úkol stanovit míru pravděpodobnosti, že se jedná o tutéž osobu.

Další technikou, kterou lze použít při ztotožňování osob, je analýza jimi napsaných textů. Zde se zkoumají podobnosti slovní zásoby, častý výskyt vybraných slovních kombinací, shodné pravopisné a typografické chyby, podobné překlepy. Z těchto informací se pak odvozuje míra pravděpodobnosti, že text byl psán toutéž osobou. Případy využití těchto funkcí zle hledat například v boji s dětskou pornografií na internetu.

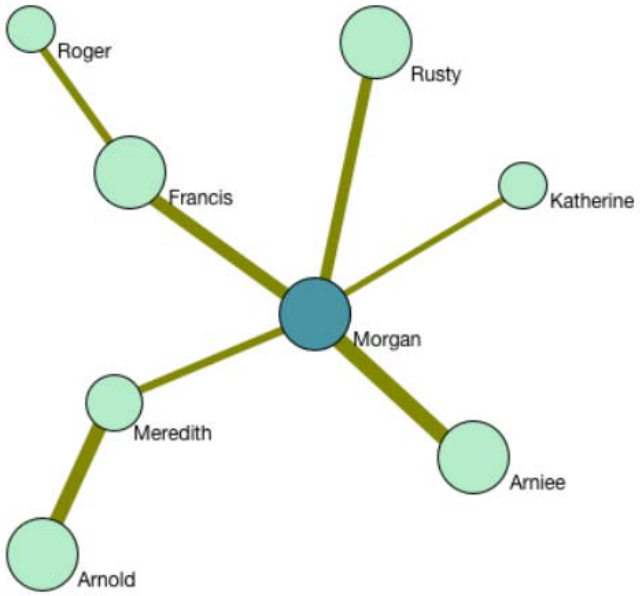
4.5 Analýza sentimentu

Analýza sentimentu se provádí pomocí vlastních nástrojů, které jsou schopny pracovat i s vícejazyčnými daty. Základním principem analýzy je hledání specifických slovních kombinací, jež indikují příslušný typ a míru sentimentu v dané doménové oblasti, jak uvádí Ministr a Ráček (2011).

Analýza emocí byla původně vyvinuta pro marketingové analýzy dat, nicméně své místo má i v oblasti policejních analýz. V tomto případě je například posuzována míra agresivity a odhodlání provést příslušný čin. Sledovány jsou i souhlasné a nesouhlasné projevy s konkrétními činy, což má své využití například při boji s extremismem.

5 Prezentace výsledků analýz

Výsledky analýz jsou prezentovány ve dvou základních formách, které doporučují Gorodov a Gubarev (2013). Buď v agregované podobě, což jsou nejčastěji grafy, nebo v podobě detailu nalezeného záznamu. Grafy jsou využívány zejména pro vizualizaci vývoje nějakého jevu v čase (čárové a sloupcové grafy), vizualizaci procentuálního zastoupení různých typů entit (koláčové grafy) nebo pro vizualizaci vztahů (síťové diagramy). V policejní praxi mají z těchto diagramů velký význam také síťové diagramy ukazující sítě zájmových osob. Ostatní typy grafů jsou vhodné spíše k vizualizaci trendů a statistik za vybraná období. Jako příklad vizualizace, která je používána zejména pro znázornění vztahů mezi osobami, jako příklad je uveden síťový graf na obr. č. 2.



Obr. 2. Vizualizace sítě zájmových osob.

Tímto způsobem jsou prezentovány výsledky, které ukazují vtahy zájmových osob. Výsledná síť zájmových osob je získána například analýzou mailové komunikace, dat ze sociálních sítí nebo diskuzí na fórech. Jednotlivé osoby jsou reprezentovány uzly. Vztah je reprezentován hranou. Šířka hrany zpravidla představuje sílu vazby, což je například intenzita komunikace. Barva hrany může ukazovat další parametry vazby, jako je například téma nebo sentiment, jak uvádí Borgatti et al. (2009).

Při vyšetřování trestných činů se hodně pracuje s detaily konkrétních nalezených entit, ať již hledaných osob nebo předmětů. Přínosem je zejména to, že se sestavuje záznam o entitě, který agreguje atributy z více zdrojů a současně s tím zobrazuje čas, kdy byly jednotlivé atributy zaznamenány, a míru jejich důvěryhodnosti. Z toho důvodu jsou velmi významnou formou výstupu i přehledy nalezených zájmových entit řazené dle relevance a detailní zobrazení entit se zvýrazněním atributů, které jsou podstatné pro danou pátrací úlohu.

6 Závěr

Architektura softwarového systému, který využívá automatické nástroje pro rozpoznávání a identifikaci osob, objektů a vztahů v nestrukturovaných datech, je postavena na využití komerčních nástrojů v datové vrstvě, ale hlavní funkcionalita takového softwarového systému je zajištěna speciálními analytickými a prezentačními funkcemi, které je třeba vytvořit na základě specifických potřeb oblasti nasazení takového systému

Softwarové řešení, jehož architektura byla v tomto příspěvku představena, je v současné době (1. polovina roku 2015) předáváno Policii ČR k pilotnímu provozu. Na tomto místě je třeba poznamenat, že se nejedná o informační systém určený koncovým uživatelům, ale jde o modulární stavebnici, z níž budou uživateli sestavovány a konfigurovány koncové aplikace, které budou pokrývat specializované případy užití dle potřeb konkrétních útvarů policie. Testovací provoz bude v roce 2015 probíhat v oblasti pátrání po odcizených uměleckých dílech. Z pohledu budoucího využití celého řešení se nabízí velká řada státních i mezinárodních institucí, ale i soukromých subjektů. Jak již bylo uvedeno, v současné době je koncovým uživatelem Policie České republiky, konkrétně útvary s celorepublikovou působností zabývající se kriminalitou. Patří sem zejména Služba kriminální policie a

vyšetřování - Policie České republiky a její útvary s celorepublikovou působností, jako jsou Útvar odhalování korupce a finanční kriminality, Útvar pro odhalování organizovaného zločinu a Národní protidrogová centrála. Na mezinárodní úrovni je řešení koncipováno tak, aby díky mezinárodním standardům a nezávislosti na jazycích bylo využitelné i Interpolem a Europolem. Robustní vícejazyčnost rovněž umožňuje využití bezpečnostními službami. Z pohledu použitelnosti v komerčních společnostech se ukazuje, že řešení je poměrně snadno nasaditelné na interní data pojišťoven, kde může napomáhat odhalování pojistných podvodů.

Výše uvedené možnosti rozšíření na další skupiny uživatelů je v současnosti předmětem dalšího výzkumu a vývoje nových modulů, jejichž úkolem je zejména integrace s okolními systémy a vývoj nových vizualizačních metod pro jednotlivé nově identifikované případy užití modulárního řešení v různých aplikačních oblastech nasazení takto zaměřených systémů.

Poděkování

Autoři článku děkují za podporu projektu VF20132015030 - Automatické rozpoznávaní a identifikace objektů v internetu a stávajících systémech PČR se zaměřením na PSEUD, dále grantu "Výzkumný tým pro modelování ekonomických a finančních procesů na VŠB - Technické univerzitě Ostrava" s referenčním číslem CZ.1.07/2.3.00/20.0296.

Seznam použitých zdrojů

Aggarwal, C. & Subbian, K. (2014). Evolutionary Network Analysis: A Survey. *ACM Computing Surveys*. 47(1), 10.1-10.36.

Ashton, T., Evangelopoulos, N. & Prybutok, V. (2014). Extending monitoring methods to textual data: a research agenda. *Quality & Quantity*. 48(4), 2277-2294.

Borgatti, S. P., Mehra, A., Brass, D. & Labianca, G. (2009). Network Analysis in the Social Sciences. *Science*. 323, 892-895.

Gorodov, E. Y. & Gubarev, V. V. (2013). Analytical Review of Data Visualization Methods in Application to Big Data. *Journal of Electrical and Computer Engineering*. 2013(22), 1-7.

Ministr, J. & Ráček, J. (2011). Analysis of Sentiment in Unstructured text. In P. Doucek & G. Chroust (Eds.), *19th Interdisciplinary Information Management Talks,* (pp. 299-304). Linz: Trauner.

Khan, F. H., Bashir, S. & Qamar, U. (2014). TOM: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, 57, 245-257.

Koppel, M., Schler, J. & Argamon, S. (2009). Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology,* 60(1), 9-26.

Sathi, A. (2012). *Big Data Analytics: Disruptive Technologies for Changing the Game*. Boise: McPress.

Scott, J. (2000). *Social Network Analysis: A Handbook*. Thousand Oaks: SAGE Publications.

Xu, J. J. & Chen, H. (2005). CrimeNet Explorer: A Framework for Criminal Network Knowledge Discovery. *ACM Transaction on Information Systems*, 23(2), 201-226.