Identifikace tématických sociálních sítí

Jiří Jelínek

Katedra managementu informací, FM, Vysoká škola ekonomická,
Jarošovská 1117/II, 377 01, Jindřichův Hradec
jelinek@fm.vse.cz

Abstrakt. V rámci tohoto projektu byl navržen postup identifikace tématických sociálních sítí pomocí analýzy výstupů webových vyhledávacích systémů po zadání určité odborné oblasti či specifického klíčového slova. Byla vytvořena metoda identifikace vlastních jmen osob a byly zkoumány možnosti automatizace této činnosti. Dále byla pozornost věnována významovému zpřesnění těchto jmen a odstranění případných duplicit. Součástí projektu bylo také vytvoření mechanizmu pro odhalování odborných vazeb mezi jedinci. Zkoumány byly i různé formy vizualizace výsledných sítí. Předkládaný příspěvek popisuje kromě výše uvedených metod i z nich vycházející praktické řešení a výsledky získané jeho testováním v praxi.

Klíčová slova: WWW, grafy, sociální sítě, NER, detekce vazeb mezi termy

1 Úvod

Hlavním motivem sepsání tohoto příspěvku je prezentovat metody a postupy použitelné v softwarovém nástroji umožňujícím, pokud možno automaticky, identifikovat a následně zobrazit odborné vazby mezi jednotlivci na základě obecně dostupných dat získaných webovými vyhledávači. Takový nástroj by mohl výrazně pomoci při orientaci „kdo je kdo“ v dané oblasti, bylo by možné identifikovat významné jedince a při odborné práci se soustředit na informace „od pramene“.
Vyhledávání vlastních jmen osob v prostředí WWW a jejich další zpracování není příliš rozšířenou službou. Je samozřejmě možné využít při hledání konkrétního jména standardní vyhledávací systémy, otázkou však je, nakolik jsou získané výsledky prakticky použitelné, jestliže na dotaz „John Smith“ dostaneme např. vyhledávačem Google 261 milionů odkazů.
Také je otázkou, zda právě takto položený dotaz nejlépe vystihuje potřeby uživatele a poskytne odpovědi, které uživatel chce získat. Většinou nás nezajímají ani tak jména samotná, jako spíše jména v určitém kontextu (oblasti), který lze pro účely vyhledávání charakterizovat vybranými klíčovými slovy. Dotaz tedy častěji směřuje spíše na zvolenou oblast, ve které nás zajímají vlastní jména osob s ní spojených a vztahy těchto osob.
Právě na základě výše uvedené úvahy byl vytvořen tento příspěvek. S postupem prací se však ukázalo, že oblast identifikace a zpracování vlastních jmen osob vyžaduje širší zkoumání a také, že výsledky mohou být užity při podpoře širšího spektra činností, než byl původní předpoklad.

Příspěvek je dále rozdělen do několika kapitol. Druhá kapitola charakterizuje současný stav v předmětné oblasti výzkumu. Třetí kapitola se věnuje navrhovaným postupům a metodám pro detekci vlastních jmen osob, zpřesnění jejich významu a detekci vazeb mezi osobami. Čtvrtá kapitola popisuje architekturu navrženého prototypu a prezentuje dosažené výsledky. Kapitola pátá shrnuje celý projekt a nastiňuje další postup. Šestá kapitola je pak závěrem příspěvku.

2 Současný stav

Celý problém detekce a následného zpracování vlastních jmen osob pro výše popsaný účel lze rozdělit do několika fází, které budou diskutovány dále.
První z nich je *detekce vlastních jmen osob*. Tato oblast je obvykle označována jako NER (Named Entity Recognition), EI (Entity Identification) či EE (Entity Extraction) a její historie sahá do 90. let minulého století. Přístupů k řešení tohoto úkolu je několik:

1. *Metody NLP* jsou jednou z prvních metod detekce vlastních jmen osob. Jejich základem je obvykle syntaktická analýza větné stavby textu a užití pravidel pro identifikaci jmen. Velmi podstatnou součástí detekčních mechanizmů je také sledování velkých počátečních písmen slov. To však může některé informační zdroje diskriminovat (některé zpravodajské agentury např. šíří zprávy psané pouze velkými písmeny). Detekovat lze nejen vlastní jména osob, ale i míst a organizací [3]. Příkladem užití gramatických pravidel může být např. tagger ANNIE [1], který je součástí balíku GATE nebo systémy FreeLing [5] či NE classifier [3].
2. Druhou možností je *statistický přístup*. Užity mohou být běžně užívané klasifikátory, zejména naivní bayesovský klasifikátor. Metoda vychází z dostatečně obsáhlé trénovací množiny. Na základě ručního ohodnocení příkladů z této množiny je při následujícím výskytu stejného termu vypočtena pravděpodobnost jeho příslušnosti k pozitivně či negativně hodnoceným příkladům.
3. Podobná metoda vychází z existence rozsáhlých *slovníků vlastních jmen osob*. Od statistického přístupu se liší především existencí pouze pozitivně hodnocených příkladů a přímým porovnáváním zkoumaného termu se slovníkem. Tento postup je popsán např. v [14]. Problémem je zde získání dostatečně obsáhlých slovníků.
4. *Využití kontextu* je rovněž zajímavý přístup k detekci vlastních jmen osob. Je založen na zkoumání bezprostředního okolí daného slova či sousloví [12], přičemž na slova blízká zkoumanému termu může být uplatněn statistický přístup, na základě kterého je pak danému termu přiřazena pravděpodobnost, s jakou se může jednat o vlastní jméno osoby.

Fáze *zpřesnění významu* je v případě vlastních jmen osob dosti komplikovaná a dosahované výsledky nejsou nikdy stoprocentní. Hlavní úkoly jsou zde následující:

1. Samostatným problémem je *čištění vstupních dat*, ve kterých mohou být gramatické chyby a přepisy. Jednou z možností je porovnávání jmen na

základě fonetické hodnotící funkce a užití rozsáhlých slovníků pro korekci chyb.

2. *Odlišit osoby se stejným vlastním jménem* – tento úkol je obvykle řešen s pomocí doplňkové informace. Tou může být např. tématická oblast, se kterou je osoba spojena, v případě autorů název jejich publikace, informace o geografické poloze, atd. Samotné odlišení (klasifikace) je pak realizováno klasifikátory pracujícími na základě strojového učení. Např. v [8] je jako doplňkový údaj použit název publikace dané osoby a pro klasifikaci je zvolen naivní bayesovský klasifikátor nebo Support Vector Machines (SVM). Tato fáze nemůže být zcela oddělena od následující.
3. Dále je nutné *identifikovat vlastní jména osob s různou formou zápisu* – řešení často vychází z čistě syntaktických pravidel definujících pro dvě formy zápisu způsob jejich porovnání a ohodnocení a upřednostňovaný výstup. Preferován může být jak co nejkratší zápis daný příjmením a případně iniciály prvního křestního jména (který je vlastně nejobecnějším označením jedince) nebo zápis co nejúplnější obsahující plná znění všech jmen.

Dalším krokem je *detekce vazeb mezi osobami* identifikovanými svými vlastními jmény. Pokud si uvědomíme, že tento proces je pouze speciálním případem detekce vazeb mezi termy, je možné při řešení vycházet právě z této širší oblasti.
Metoda navržená v [11] např. detekuje vazby termů na základě jejich současného výskytu v dokumentech. Ohodnocení vazeb a uplatnění prořezávacích technik je založeno na podmíněných pravděpodobnostech jejich výskytu (každá vazba je chápána jako orientovaná a je tedy ohodnocena v obou směrech). V [11] jsou uvedeny i další navazující postupy možného využití takto získaných dat o vztazích termů.

3 Navržené postupy

Při návrhu metod detekce a zpracování vlastních jmen osob bylo hlavním cílem definovat kompletní metodiku celého procesu tak, aby na jejím základě mohla být vytvořena použitelná aplikace. Celý postup byl rozdělen do následujících fází:

1. Detekce jmen osob
2. Zpřesnění významu (identifikace)
3. Detekce vazeb mezi jmény

Jednotlivé fáze budou nyní probrány podrobněji.

3.1 Detekce vlastních jmen osob

Úkol, který je nutné v této části vyřešit, lze definovat takto: mějme zadaný prostý text obsahující vlastní jména osob, požadovaným výstupem je seznam těchto jmen.
V popisovaném případě je vstupem WWW stránka, jejíž URL je buď přímo zadané nebo získané jako součást výstupu vyhledávače. Samotná detekce jmen probíhá v několika dále popsaných fázích.
První z nich je *použití masky na vstupní text*. Tento krok odhaluje možné kandidáty na vlastní jména osob. Přípustné formy zápisu jsou v zásadě dvě: „jméno1 jméno2

příjmení“ nebo „příjmení, jméno1 jméno2“. Na pozicích křestních jmen mohou být rovněž pouze iniciály, druhé křestní jméno může být vynecháno.

Paralelně s tímto způsobem detekce probíhá *identifikace NLP* s pomocí balíku „Named Entity Tagger“ [3], jehož výstup je sloučen s výstupy výše uvedené metody.

Kandidáti z takto získané množiny jsou následně ohodnoceni několika různými technikami. Cílem hodnocení je kvantifikovat šanci, s jakou je kandidát skutečně vlastním jménem osoby (čím vyšší kladné hodnocení, tím větší šance, že jde o vlastní jméno osoby).

První část ohodnocení vychází z *kontroly křestních jmen*. Pro tento krok byla ze serveru [2] extrahována běžně používaná křestní jména pro celou škálu jazyků (angličtina, němčina, čeština, arabština, čínština, atd.). Dalším zdrojem referenčních dat byla databáze „DataBase systems and Logic Programming” (DBLP) [4] obsahující bibliografické informace o obsahu hlavních časopisů a sborníků zaměřených na výše uvedenou oblast. Vytvořená databáze (cca 60 000 unikátních jmen) je pak užita ke kontrole křestních jmen, nalezení kandidáta v databázi vede ke zvýšení jeho kladného ohodnocení o hodnotu k_f.

Stejný postup je uplatněn rovněž při *kontrole příjmení*. Celý systém je zaměřen na angličtinu, proto byl za základ referenční databáze příjmení použit výstup sčítání obyvatel USA, kde jsou nejčastější příjmení uvedena [6]. Tento zdroj byl dále doplněn z [9], kde jsou uvedena příjmení studentů amerických univerzit z roku 2003 a z DBLP [4]. Získaná databáze obsahuje cca 217 000 příjmení a identifikace kandidáta zvýšila jeho ohodnocení o koeficient k_l.

Další formou ohodnocení je *využití databáze podstatných jmen z projektu WordNet* [15] obsahující cca 143 000 unikátních položek. Ty jsou porovnávány s příjmeními kandidátů. V případě, že příjmení se nevyskytuje ve WordNetu, je zvýšeno pozitivní hodnocení kandidáta o koeficient k_w. Tato kontrola je založena na úvaze, že slova bez reálného významu mohou být příjmeními.

Další metody ohodnocení jsou založeny na *statistickém principu učení z předchozích rozhodnutí*. Systém uchovává jak pozitivně klasifikované kandidáty, ze kterých se stávají regulérní termy, tak i negativně klasifikované případy. Každý nový kandidát je ohodnocen na základě výpočtu koeficientu

$$k_s = k_{sm} \frac{c_p - c_n}{c_p + c_n}, \tag{1}$$

kde c_p je počet pozitivně hodnocených výskytů daného jména, c_n počet negativně hodnocených případů a k_{sm} je volitelný koeficient odrážející váhu tohoto hodnotícího kritéria. Tento systém hodnocení lze samostatně uplatnit jak na příjmení, tak na křestní jména.

Poslední kritérium vychází z modelu popsaného v [12] a postupu uvedeného v předchozím odstavci. Hodnocení kandidáta je zvýšeno o hodnotu k_c podle *výskytu slov v jeho bezprostředním okolí*, které sahá 3 slova před a tři slova za příslušného kandidáta.

Výstupem fáze detekce je seznam kandidátů, u nichž je pro výpočet jejich výsledného ohodnocení použit následující vzorec:

$$h = k_f + k_l + k_w + k_s + k_c \tag{2}$$

Vyjádření vah jednotlivých členů je dáno již samotnou volitelnou hodnotou jednotlivých koeficientů (výjimkou jsou k_s a k_c , jejichž váhy jsou dány maximálními hodnotami k_{sm} a k_{cm}). Výsledné ohodnocení by jistě bylo možné vyjádřit i jiným vztahem, výběr optimálního výpočtu a nastavení koeficientů mohou být předmětem dalšího výzkumu.
Seznam kandidátů může být následně prezentován uživateli k ruční klasifikaci nebo ohodnocen automaticky. První možnost je podstatná zejména v počátečních fázích, kdy není k dispozici dostatek klasifikovaných příkladů vlastních jmen. Později již lze využít klasifikaci na základě uživatelem zadaných mezních hodnot hodnocení h_{min} a h_{max}.
V případě $h > h_{max}$ je kandidát považován za vlastní jméno, pokud $h < h_{min}$, je jeho výsledné hodnocení negativní. S případy, kdy platí $h_{min} < h < h_{max}$, lze naložit různě. Vhodnou cestou se zdá být jejich vymazání ze seznamu kandidátů nebo jejich ruční hodnocení.
Výstupem celé části detekce vlastních jmen osob je tedy seznam klasifikovaných kandidátů, z nichž jsou dále zpracovávány pouze pozitivní případy, (detekovaná vlastní jména).

3.2 Zpřesnění významu

Metoda zpřesnění významu se zaměřuje především na identifikaci osob a selekci jediné formy zápisu vlastního jména pro danou osobu, přičemž oba úkoly jsou řešeny současně.
Nejprve jsou porovnávány různé formy zápisu vlastních jmen a je testováno, zda označují stejnou osobu. Za kritérium shody je bráno stejné příjmení a shoda křestních jmen (prvních) nebo jejich iniciálů. Z takto zjištěných možných zápisů jednoho jména je vybrán ten, který je nejúplnější (pokud možno plné znění všech jmen).
Problém identifikace osoby je zjednodušeně řešen s pomocí doplňkové informace, kterou tvoří *téma* (charakterizované klíčovým slovem nebo slovy), ke kterému má daná osoba vztah (po jehož zadání do vyhledávače bylo dané jméno získáno). Při výběru preferované formy zápisu je proto tento údaj brán v úvahu a porovnávány jsou jen termy z jedné tématické oblasti. Předpokladem tohoto řešení je, že v dané oblasti se vyskytuje pouze jedna osoba s jedinečnou kombinací jméno - příjmení.

3.3 Identifikace souvislostí mezi termy

Identifikace souvislostí mezi termy je prováděna na základě výskytu těchto termů společně v jednotlivých vstupních dokumentech (WWW stránkách). Použitý algoritmus vychází z postupu uvedeného v [11] s drobnými úpravami.
Dále uvedené výpočty jsou vždy vztaženy k množině dokumentů S vzniklé sjednocením WWW stránek z tématických skupin definovaných výrazy zadanými pro jejich vyhledání do vyhledávače Google. Váha konkrétních termů se tak může lišit podle parametrů vyhledávání a je definována jako

$$w_{iS} = \frac{\sum_{k=1}^{K_S} c_{ik}}{\sum_{k=1}^{K_S} n_k}, \tag{3}$$

kde w_{iS} je váha termu t_i vzhledem k množině tématických skupin S, K_S počet tématických skupin sjednocených v S, n_k počet dokumentů v dané tématické skupině a c_{ik} počet dokumentů s termem t_i v tématické skupině k.

Podle [11] tvoří termy s $w_{iS} > práh$ množinu významných termů V, která slouží za základ dalšímu postupu. Na té jsou dále definovány dvojice termů (t_i , t_j). Pro každou takovou dvojici a pro množinu S lze vypočítat výraz

$$p_{ijS} = \frac{2\sum_{k=1}^{K_S} c_{ijk}}{\sum_{k=1}^{K_S} (c_{ik} + c_{jk})}, \quad i \neq j \tag{4}$$

kde K_S je počet tématických skupin sjednocených v S, c_{ijk} počet dokumentů se současným výskytem termů t_i i t_j v tématické skupině k a c_{ik} , resp. c_{jk} jsou počty dokumentů ve skupině k, kde se vyskytuje term t_i , resp. t_j.

Pro stanovení významnosti vazby mezi termy t_i a t_j byla zvolena hodnota

$$h_{ijS} = k(w_{iS} + w_{jS}) + (1-k)p_{ijS} \tag{5}$$

Tato hodnota charakterizuje význam vazby mezi osobami s vlastními jmény t_i a t_j v dané množině tématických skupin S. Vztah je založen na síle vazby p_{ijS} , vycházející ze společného výskytu termů t_i a t_j, a na významnosti uvedených termů. Volitelný koeficient k z intervalu <0,1> umožňuje zdůraznit složku vycházející z významnosti termů ($k \to 1$) nebo složku založenou na ohodnocení dané vazby ($k \to 0$). Pro zařazení vazby do výstupu musí být $h_{ijS} > m$, kde m je uživatelem definovaná mezní hodnota.

Výsledkem této fáze je seznam dvojic termů, které se vyskytují společně včetně ohodnocení jejich vazby hodnotou h_{ijS}.

K vizualizaci dat získaných výše uvedenými postupy je použita knihovna Graphviz [7]. Z dostupných algoritmů pro tvorbu rozvržení grafu byl vybrán algoritmus NEATO. Základním výstupem vizualizace je zobrazování termů a jejich vazeb z vybraných tématických oblastí. Vzdálenost jednotlivých uzlů byla nastavena úměrně hodnotě $1/ h_{ijS}$, barva uzlů podle váhy termů w_{iS}.

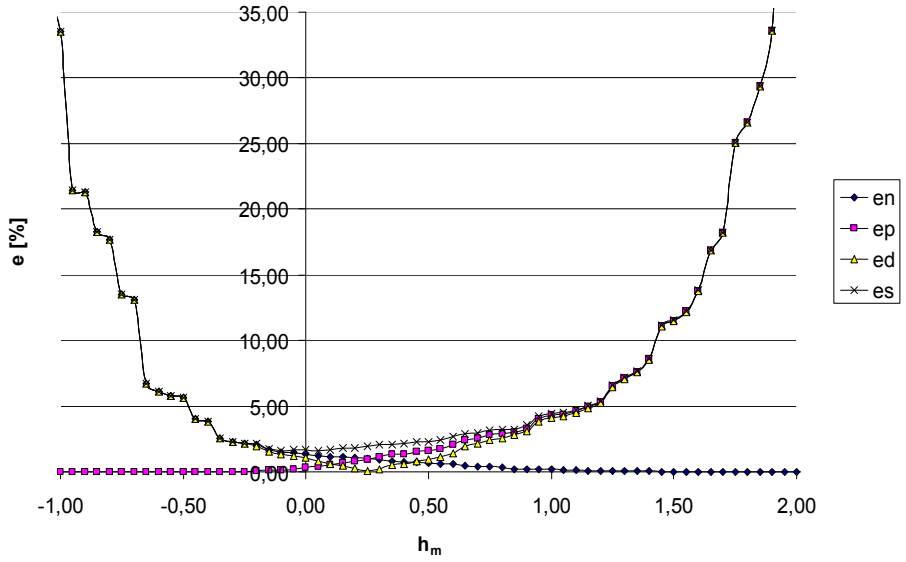
4 Architektura prototypu a dosažené výsledky

Výše uvedené postupy byly implementovány do prototypu webové aplikace napsané v PHP a MySQL. Ta umožňuje realizovat všechny uvedené činnosti:

1. Načtení výstupů vyhledávače Google pro zadané téma a s volitelným počtem odkazů ve výstupu – vyhledávač Google je zde použit pro nalezení relevantních stránek k danému tématu. Volit lze rovněž mezi dvěma způsoby

vyhledávání (fráze či seznam slov). Nalezené stránky systém následně načítá a detekuje v nich kandidáty na vlastní jména osob.

2. Načtení zadaného URL a jeho přidání k zadanému tématu – stejná činnost jako v předchozím bodu, ale stránka není vyhledávána, nýbrž zadána přímo.
3. Automatické ohodnocení nalezených kandidátů postupy uvedenými výše v tomto příspěvku.
4. Volitelné zobrazení kandidátů pro klasifikaci uživatelem – je prezentován seznam kandidátů s jejich ohodnocením a výstupem přednastaveným podle kritérií h_{min} a h_{max}. Tato činnost není prováděna v automatickém režimu.
5. Prezentace seznamu dosud analyzovaných témat s počtem analyzovaných stránek ke každému z nich.
6. Analýza vstupů a následné zobrazení síťového grafu souvislostí mezi termy – kritériem pro zařazení dané vazby je $h_{ijS} > m$ (viz výše). Jednotlivé uzly (osoby) mohou být barevně odlišené podle hodnot w_{iS}.
7. Grafické zobrazení vazeb vybraného jedince podle příjmení – toho lze zvolit kliknutím na předchozí graf tématické oblasti. V případě shody příjmení jsou vypsány všechny vyhovující termy.



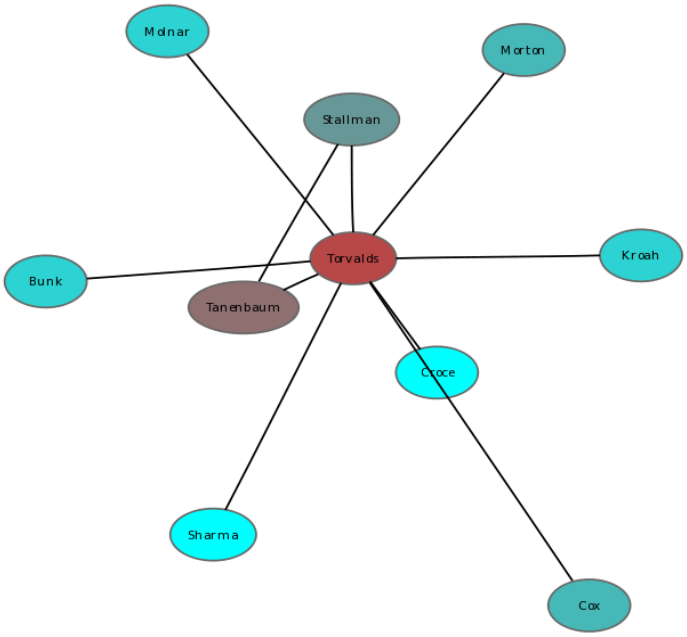
Obr. 1. Graf závislosti chyb e_d, e_p, e_n a e_s na hodnotách h_m

Aby aplikace umožňovala rovněž automatizovaný režim provozu, kdy uživatel zadá seznam požadovaných témat k analýze a systém je zpracuje, je nutné stanovit uživatelsky zadané koeficienty. Ve fázi detekce vlastních jmen jde zejména o hodnoty h_{min} a h_{max}. Podle jejich volby jsou pak hodnoceni jednotliví kandidáti, přičemž jejich vlastní hodnocení h je závislé na koeficientech k pro jednotlivé metody hodnocení.
Volba k_f, k_l, k_w, k_{sm}, k_{cm} musí být provedena tak, aby pozitivně a negativně klasifikovaní kandidáti měli maximálně odlišné hodnoty h při minimální chybě

klasifikace. Pro zjednodušení bylo při experimentech stanoveno $h_{min} = h_{max}$, takže bylo nutné nalézt pouze jedinou mezní hodnotu h_m.
Pro nalezení její optimální velikosti na základě již klasifikovaných termů byly pro h_m z vybraného intervalu vypočteny hodnoty chyb

$$e_n = \frac{c_{n+}}{c_n} , \; e_p = \frac{c_{p-}}{c_p} , \qquad (6)$$

kde c_{n+} je počet negativně hodnocených příkladů s $h > h_m$, c_n celkový počet negativně hodnocených příkladů, c_{p-} je počet pozitivně hodnocených příkladů s $h < h_m$ a c_p celkový počet pozitivně hodnocených příkladů. h_m může být stanovena různým způsobem. Jednou z variant je stanovení h_m podle minimální hodnoty výrazu $e_d = | e_p - e_n |$. Tento postup zaručuje stejnou velikost chyb e_p a e_n . Jinou možností je např. použití kritéria minimální souhrnné chyby $e_s = e_p + e_n$. Výběr nejvhodnější metody stanovení h_m bude předmětem dalšího výzkumu.
Grafy průběhů chyb pro vybrané nastavení koeficientů k_f , k_l , k_w , k_{sm} , k_{cm} jsou uvedeny na obr.1. Graf naznačuje, že vhodným nastavením může být h_m = 0,25 při volbě podle hodnot e_d nebo h_m = -0,1 při volbě podle e_s .
S popsaným systémem byly provedeny první experimenty. Systém nyní obsahuje data z 397 analyzovaných WWW stránek příslušných k 17 různým tématům. Počet klasifikovaných termů dosáhl 16671, z toho 3079 pozitivně hodnocených.
Z dosažených výsledků vyplývá, že použité metody jsou zcela životaschopné a mohou sloužit jako základ pro další výzkum a vývoj. Jako příklad výstupu je na obr. 2 uveden graf identifikované tématické sítě pro téma „linux“ založené na analýze 51 webových stránek.



Obr. 2. Identifikovaná tématická síť pro téma „linux“

Barva termů na obrazovce (odstín v tisku) odpovídá jejich váze, délky hran byly v zadání grafu voleny nepřímo úměrně síle příslušné vazby, což vykreslovací algoritmus respektoval v rámci možností 2D zobrazení.

5 Shrnutí a další postup

Na základě provedených experimentů lze konstatovat, že aplikace vystavěná na zde popsaných metodách identifikace tématických sociálních sítí poskytuje zajímavé výsledky a vytváří vhodný základ pro další výzkum v této oblasti.
Další postup se soustředí na některá vylepšení, která by mohla možná užití aplikace dále rozšířit a zpřesnit dosahované výsledky. Mezi ně patří například rozšíření vstupních importních filtrů. V současné době je primárním zdrojem dat vyhledávač Google, a to především z důvodu zajištění obecného zdroje informací. Počítá se však s implementací dalších importních filtrů pro speciální data. Jako nejzajímavější se jeví užití dat z citačních serverů.
V oblasti detekce vlastních jmen by bylo vhodné se zamyslet nad postupy stanovení vhodných hodnot k_f , k_l , k_w , k_{sm} , k_{cm} . Protože jde v podstatě o optimalizační úlohu, možnou cestou by zde bylo užití genetických algoritmů.
Na základě dalších experimentů bude též nutné se dále zabývat metodikou automatizovaného stanovení mezních hodnot h_{min} a h_{max} (event. h_m).
Pro zpřesnění významu nalezených termů by bylo užitečné uvažovat o vývoji a implementaci dokonalejších technik založených na pokročilých postupech analýzy získaných dat.
Pro zobrazení výstupů by mohlo být užito 3D zobrazení pomocí jazyka VRML, které je popsáno např. v [10].

6 Závěr

Popsané teoretické metody informačními technologiemi podporované identifikace tématických sítí jsou dalším příspěvkem do aktuální oblasti tvorby a detekce potenciálních sociálních sítí. Navrhovaný prototyp aplikace je pak ukázkou jejich konkrétního uplatnění a získané výstupy mohou být přímo užity v praxi a to nejen v oblasti vědeckého výzkumu, ale všude tam, kde je potřeba identifikovat tématicky definované sociální sítě (např. v oblasti finančnictví, v kriminalistice, ekonomice, atd.).
Postup byl prakticky otestován, přičemž objevil významné vazby v daných tématických oblastech. V současné době probíhají další experimenty zaměřené především na výzkum modifikací a rozšíření uvedené metodiky.

Reference

1. Annie Named Entity Tagger, In: *http://www.media-style.com/index.jsp?folderPK=754*, October 2007
2. Behind the Name - the Etymology and History of First Names, In: *http://www.behindthename.com/*, October 2007
3. CCG: Software – Named Entity Tagger. In: *http://l2r.cs.uiuc.edu/~cogcomp/asoftware.php?skey=NE#tools*, October 2007
4. DBLP Bibliography, *In: http://dblp.uni-trier.de/xml/*, October 2007
5. FreeLing Home Page. In: *http://garraf.epsevg.upc.es/freeling/*, October 2007
6. Frequently Occurring Names from the 1990 Census, In: *http://www.census.gov/genealogy/www/freqnames.html, October 2007*
7. Graphviz, In: *http://www.graphviz.org/*, October 2007.
8. Han, H.; Giles, L.; Zha, H.; Li, C.; Tsioutsiouliklis, K.: Two supervised learning approaches for name disambiguation in author citations. In: *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries* (Tuscon, June 2004). JCDL '04. ACM Press, New York, 2004, pp. 296-305.
9. ICU Project at the Data Privacy Laboratory, In: *http://privacy.cs.cmu.edu/dataprivacy/projects/icu/datainfo.html*, October 2007
10. Jelínek J.; Kunčar D.; Přibil J.: Vizualizace textových dat pomocí grafů, *Konference Znalosti 2006*, Hradec Králové, únor 2006, In: Paralič J., Dvorský J., Krátký M. (eds.): Znalosti 2006, pp. 276-279, ISBN 80-248-1001-8, VŠB-Technická univerzita Ostrava, Fakulta elektrotechniky a informatiky, 2006
11. Jelínek, J.: Využití vazeb mezi termy pro podporu uživatele WWW. *Mezinárodní konference Znalosti 2005, 9. – 11. 2. 2005, Stará Lesná, Slovensko,* In: Sborník příspěvků 4. ročníku konference Znalosti 2005, pp. 218-225, VŠB-TUO FEI Ostrava, ISBN: 80-248-0755-6
12. Minkov, Einat; Wang, Richard; Cohen, William: Extracting Personal Names from Email: Applying Named Entity Recognition to Informal Text. In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, October 2005, Association for Computational Linguistics
13. Named entity recognition - Wikipedia, the free encyclopedia, In: *http://en.wikipedia.org/wiki/Named_entity_recognition*, October 2007
14. Stevenson, M.; Gaizauskas, R.: Using corpus-derived name lists for named entity recognition. *In: Proc. of ANLP*, Seattle, 2000.
15. WordNet, In: *http://www.cogsci.princeton.edu/~wn/*, October 2007.
16. Xia, Jingfeng: Personal name identification in the practices of digital repositories. In: *Program: Electronic Library & Information Systems*, 2006, 40(3): pp. 256-267

Annotation:

Identification of the Thematic Social Nets

In the scope of this project we proposed a procedure of identification of the thematic social nets by means of the analysis of output from web search engines after entering a specific specialized area or a specific keyword. A method of identification of personal names was constructed and the possibilities of automation of this activity were tested. Then the attention was devoted to the sense recognition of these names and deleting of unwanted duplicates. Another part of the project was the creation of a mechanism for detection of specific relationships amongst the individuals. Different forms of visualization of these networks were also studied. The proposed article describes together with the above stated methods the practical solution based on them and results obtained from its testing in use.