Odhalování tématických sociálních sítí s pomocí WWW

Jiří JELÍNEK
Vysoká škola ekonomická v Praze
jelinek@fm.vse.cz

INFORUM 2008: 14. konference o profesionálních informačních zdrojích
Praha, 28. - 30.5. 2008

Abstrakt.

Sociální sítě a jejich detekce a tvorba patří k velmi aktuálním tématům současnosti. Jejich praktické užití je velmi rozmanité, jednou z mnoha možností je i identifikace významných jedinců a především úzce spolupracujících odborných týmů v různých oblastech vědy a výzkumu.

Pro odhalení tématické sociální sítě je potřeba získat dostatek vstupních dat, přičemž jako ideální a obsáhlý zdroj se zde nabízí WWW. Efektivní využití informací dostupných na WWW však není zcela jednoduché. Již identifikace vlastních jmen osob je dosti komplikovaná. Podobně obtížné je jednoznačně identifikovat osoby a vyhledat jejich vzájemné vazby či vztahy, na základě kterých pak lze vytvořit příslušnou sociální síť. Použitelnost získaných výstupů je však značná. Mohly by výrazně pomoci při orientaci „kdo je kdo“ v dané oblasti a při odborné práci by umožnily soustředit se na informace „od pramene“.

V rámci příspěvku je prezentován jak současný stav v oblasti detekce tématických sociálních sítí, tak zejména postupy identifikace těchto sítí s využitím analýzy výstupů webových vyhledávacích systémů. Hlavním cílem je ukázat metody prakticky použitelné pro pokud možno automatickou identifikaci a následnou vizualizaci odborných vazeb mezi jednotlivci. Kromě výše uvedených metod je prezentováno i z nich vycházející praktické řešení a výsledky získané jeho testováním v praxi.

1 Úvod

Představme si následující situaci. Jsme postaveni před úkol zorientovat se v zadané oblasti. K tomu nám mohou pomoci informace o osobách v této oblasti působících a jejich vzájemných vazbách. Nemusí vždy jít o oblast vědeckou nebo odbornou, zajímavé je též mít přehled např. o vzájemných vazbách osob v oblasti obchodu, politiky, atd. Jak ale zjistit, kdo s kým spolupracuje, kdo je podstatný a kdo nikoliv? K tomu by se nám velmi hodil nástroj schopný identifikovat významné jedince a jejich vztahy. S jeho pomocí bychom tedy mohli odhalit tzv. tématickou sociální síť. Sociální síť proto, že bude zachycovat především sociální vztahy mezi jedinci, a tématickou, neboť se omezíme na vybranou oblast a jako podklady použijeme informace o tématicky podložených vazbách.

Jako možný zdroj informací se zde nabízí Internet, avšak získání uvedených údajů není triviální záležitost. S postupem prací se ukázalo, že oblast identifikace tématických sociálních sítí a zpracování vlastních jmen osob vyžaduje širší zkoumání a také, že výsledky mohou být užity ve více oblastech, než byl původní předpoklad. Tento příspěvek se pokusí ukázat složitost celé problematiky.

2 Současný stav

2.1 Dostupná řešení

Uvedenou úlohu je možné s využitím informací z Internetu řešit různými způsoby. Ten nejjednodušší by bezpochyby spočíval v užití existující internetové služby, která by nám potřebné informace našla a prezentovala. Jako vstup by sloužila definice oblasti našeho zájmu, výstupem by pak byla výše charakterizovaná tématická sociální síť. Bohužel brzy zjistíme, že taková služba není pro libovolně definovanou oblast k dispozici.

Existují však dílčí řešení. Základem pro použití většiny z nich je nutná definice naší zájmové oblasti, tedy určitého kontextu, v rámci kterého se chceme pohybovat. Výjimkou je pouze užití takových zdrojů informací, které jsou již tématicky zaměřené a u kterých není nutné další upřesnění. V budoucnu se jistě uplatní sémantický popis, v současnosti však bude běžnější charakteristika pomocí klíčového slova nebo slov.

Jestliže jsme schopni taková slova stanovit, můžeme využít některou službu založenou na vyhledávání pomocí klíčových slov. Mezi takové služby patří všechny obecné vyhledávače na WWW, např. Google [8]. Tyto služby však nejsou optimalizovány na identifikaci osob a jejich vazeb, ale spíše na obsah a popis tématu. Na druhou stranu však tímto způsobem můžeme nalézt největší množství zdrojů a to prakticky z jakékoliv oblasti.

Můžeme se též zaměřit na specializované služby, jejichž výstup standardně zahrnuje identifikaci osob. Příkladem může být např. Google Scholar [9] umožňující k danému informačnímu pramenu získat rovněž identifikaci autora. Do této kategorie by patřilo rovněž vyhledávání v citačních indexech, mezi něž patří např. CiteSeer [4]. Specifickým, ale zajímavým zdrojem z českého WWW prostoru by mohl být např. obchodní rejstřík na serveru justice.cz. Bohužel u těchto zdrojů bývá použit výstupní formát, který pro detekci osob vyžaduje další netriviální zpracování. Jinak tomu je např. u ISI Web of Knowledge [20], pokud využijete placenou službu a export výsledků do některého ze strukturovaných formátů (např. MS Excel), kde lze identifikaci osob realizovat podstatně snadněji. Zajímavým zdrojem poskytujícím informace o významných a často citovaných jedincích je též volně dostupná služba ISI Highly Cited [13].

Žádná z uvedených služeb však stále nedokáže efektivně pracovat s vazbami mezi nalezenými jedinci a zobrazovat tématickou sociální síť. V tomto směru postoupila nejdále patrně DBLP, databáze článků z oblasti DataBase systems and Logic Programming [6] a především prohlížeč tohoto zdroje [5], který již některé funkce obsahuje.

Nevýhodou posledně jmenovaných systémů je jejich zaměření na určitou oblast. Přitom prostor WWW nabízí značné množství dat, která jsou však bohužel často uschována způsobem krajně nevhodným pro další strojové zpracování. Proto je při odhalování tematických sítí s pomocí tohoto prostoru nutno postupovat od samého počátku.

2.2 Identifikace tématických sociálních sítí

Podívejme se tedy podrobněji, jaké kroky musíme vyřešit, abychom byli schopni odhalit tématickou sociální síť:

1. Interakce s WWW prostorem
2. Detekce vlastních jmen osob
3. Identifikace osob
4. Detekce vazeb mezi osobami
5. Vizualizace

Prvním krokem je bezpochyby ***interakce s WWW prostorem***. Tento úkol lze dále rozdělit na dvě hlavní fáze, z nichž první je *vyhledání odkazů na relevantní informační zdroje*. Vhodným nástrojem zde jsou především klasické webové vyhledávače jako je např. Google. Jak již bylo uvedeno výše, nemusí však jít o zdroj jediný. Použitelné jsou rovněž služby zaměřené na specifická data (Citeseer). Základem úspěchu je vždy schopnost dobře definovat námi požadovanou tématickou oblast nejčastěji pomocí klíčových slov, případně vytvoření vhodného rozhraní pro komunikaci s vybranou službou.

Po vyhledání relevantních odkazů je ještě nutné provést *načtení příslušných stránek* z WWW prostoru pro další zpracování.

Podle [14] je následujícím krokem ***detekce vlastních jmen osob***. Tato činnost spadá do oblasti označované jako NER (Named Entity Recognition), EI (Entity Identification) či EE (Entity Extraction). Smyslem je odhalit v blíže nespecifikovaném textu vlastní jména, přičemž v našem případě se omezíme na vlastní jména osob. Přístupů k řešení tohoto úkolu je několik.

Často používané pro detekci vlastních jmen osob jsou *metody NLP (Natural Language Processing)*. Jejich základem je obvykle analýza větné stavby textu a užití zadaných pravidel pro identifikaci jmen. Podstatnou roli mezi těmito pravidly hraje sledování velkých počátečních písmen slov. Takové pravidlo však nemusí být vždy uplatnitelné (některé zpravodajské agentury např. šíří zprávy psané pouze velkými písmeny). Příkladem užití gramatických pravidel může být např. tagger ANNIE [1], který je součástí balíku GATE nebo systém NE classifier [3].

Druhou cestou je *statistický přístup*. Metoda vychází z dostatečně obsáhlé trénovací množiny, tedy souboru příkladů slov či spojení (termů), u kterých již bylo jinou cestou rozhodnuto, zda vyjadřují vlastní jména osob (pozitivní či negativní klasifikace). Při následujícím výskytu stejného termu je vypočtena pravděpodobnost jeho příslušnosti k pozitivně či negativně hodnoceným příkladům.

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 [18]. Problémem je zde získání dostatečně obsáhlých slovníků jmen.

Zajímavým postupem je *využití kontextu*, ve kterém se daný term vyskytuje. Postup je založen na předpokladu, že existují určitá slovní spojení, před kterými nebo po kterých s velkou pravděpodobností následuje vlastní jméno osoby. Je tedy zkoumáno bezprostřední okolí daného slova či sousloví [16], přičemž na toto okolí je uplatněn výše uvedený statistický přístup, na základě kterého je pak danému termu přiřazena pravděpodobnost, s jakou se může jednat o vlastní jméno osoby.

Výstupem druhého kroku zpracování je identifikace vlastních jmen osob v textu nebo textech. Tato jména však mohou být napsána různým způsobem, mohou označovat různé jedince, atd. Proto následuje fáze ***identifikace osob***. Ta je v případě vlastních jmen osob dosti komplikovaná a dosahované výsledky nejsou nikdy stoprocentní. Co je zde potřeba vyřešit?

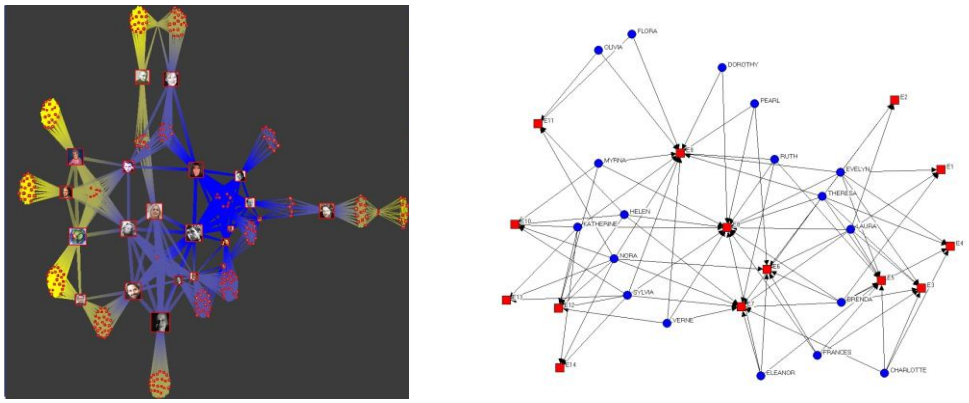
Především jde o *čištění vstupních dat* od gramatických chyb a přepisů. Jednou z možností je porovnávání jmen na základě jejich výslovnosti.

Dále je nutné *sjednotit formu zápisu jmen*. Vlastní jména mohou být zapsána různým způsobem a přesto označovat stejnou osobu. Typické je použití zkratek místo křestních jmen, vynechání druhého křestního jména nebo různé pořadí zápisu jmen. Řešení často vychází pouze z pravidel definujících pro dvě formy zápisu způsob jejich porovnání a ohodnocení. Pokud je zjištěna vysoká shoda, jsou obě formy upraveny na upřednostňovaný výstup. Preferován může být jak co nejkratší zápis daný příjmením a případně iniciály prvního křestního jména, který je však velmi obecný a snadno zaměnitelný, nebo zápis co nejúplnější obsahující plná znění všech jmen v definovaném pořadí.

Rovněž je potřeba *odlišit osoby se stejným vlastním jménem*. Tento úkol je obvykle řešen s pomocí doplňkové informace, která nám umožní takové dvě osoby od sebe odlišit. Takovou informací může být např. tématická oblast, se kterou je osoba spojena, v případě článku např. použitá klíčová slova. Samotné odlišení je pak realizováno klasifikátory pracujícími na základě strojového učení. Např. v [11] je jako doplňkový údaj použit název publikace dané osoby. Tento a předchozí krok jsou obvykle vzájemně provázány a nelze je jednoduše oddělit.

Další fází nutnou pro identifikaci tématické sítě je ***detekce vazeb mezi osobami*** identifikovanými svými vlastními jmény. Jedná se zde o speciální případ odhalování spojitostí mezi libovolnými slovy nebo spojeními (termy), což je úkol spadající do oblasti označované jako Text Mining. Lze zde tedy použít postupy definované v této sféře. Např. metoda navržená v [15] detekuje vazby termů na základě jejich současného výskytu v dokumentech. Vyskytují-li se dva termy často společně v dokumentu, lze očekávat, že mezi nimi existuje určitá spojitost. Ohodnocení vazeb a další použité postupy jsou založeny na analýze počtů společných výskytů a významnosti termů. V [15] jsou uvedeny i další navazující postupy možného využití takto získaných dat o vztazích termů.

Poslední fází je ***vizualizace výstupů***. Již klasicky se používají techniky založené na grafech, ze kterých sociální sítě vycházejí. Je možné použít některý z programů pro zobrazování grafů jako je např. Tulip [19] nebo NetDraw [17]. Tyto nástroje umožňují rovněž různé jednoduché analytické činnosti nad zobrazovanými sítěmi.



Obr. 1. Příklady zobrazení sociálních sítí aplikacemi Tulip a NetDraw
(Zdroj: http://tulip.labri.fr/sample04.php, http://www.analytictech.com/Netdraw/netdrawsamples.htm)

Jinou variantou je použít pouze zobrazovací knihovnu, např. Graphviz [10], poskytující pouze vykreslení sítě.

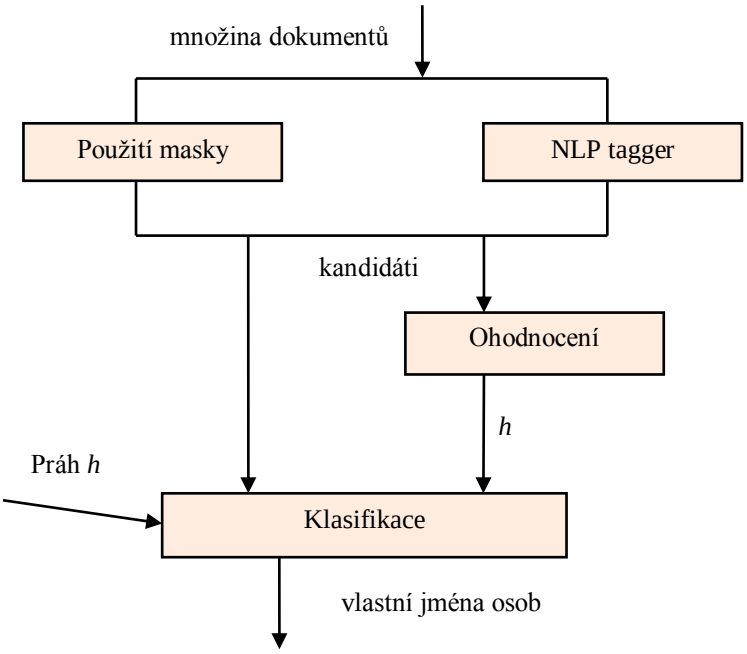
3 Navržené postupy a pilotní aplikace

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 pilotní aplikace. Popisované postupy vychází z [14], kde je také možné získat podrobnější informace.

3.1 Detekce vlastních jmen osob

Úkol, který je nutné v této části vyřešit, lze definovat takto: v zadaném textu nalézt 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. Jak je vidět z obr. 2, samotná detekce jmen probíhá v několika fázích [14].



Obr. 2. Schéma zpracování dat ve fázi detekce vlastních jmen osob

První z těchto fází je použití masky na vstupní text. Tento krok odhaluje možné kandidáty na vlastní jména osob podle přípustné formy zápisu. Paralelně s tímto způsobem detekce probíhá identifikace NLP s pomocí balíku „Named Entity Tagger“ [3], jehož výstup je sloučen s výstupy maskování.

Kandidáti z takto získané množiny jsou následně ohodnoceni několika různými technikami, přičemž cílem je kvantifikovat šanci, s jakou je kandidát skutečně vlastním jménem osoby (čím vyšší 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) [6] obsahující bibliografické informace o obsahu hlavních časopisů a sborníků zaměřených na výše uvedenou oblast. Vytvořená databáze (cca 61 000 unikátních jmen) je pak užita ke kontrole křestních jmen.

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

[7]. Tento zdroj byl dále doplněn z [12], kde jsou uvedena příjmení studentů amerických univerzit z roku 2003 a z DBLP [6]. Získaná databáze obsahuje cca 219 000 příjmení.

Další formou ohodnocení je využití databáze podstatných jmen z projektu WordNet [21] obsahující cca 143 000 unikátních položek. Ty jsou porovnávány s možnými příjmeními. Tato kontrola je založena na úvaze, že slova bez reálného významu mohou být příjmeními. Proto pokud se slovo nevyskytuje ve WordNetu, je větší šance, že půjde o příjmení.

Další metody ohodnocení jsou založeny na statistickém principu učení z předchozích rozhodnutí. Systém uchovává jak pozitivně klasifikované kandidáty na vlastní jména, tak i negativně klasifikované případy. Každý nový kandidát je ohodnocen na základě převažujícího hodnocení pro tento term v minulosti. Tento systém hodnocení lze uplatnit jak na příjmení, tak na křestní jména.

Poslední kritérium vychází z modelu popsaného v [16] a postupu uvedeného v předchozím odstavci. Hodnocení kandidáta je zvýšeno podle slov z kontextu (bezprostředního okolí), které sahá tři slova před a tři slova za příslušného kandidáta.

Výstupem fáze detekce je ohodnocený seznam kandidátů. Ten 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 automatickou klasifikaci, která samostatně stanoví mezní ohodnocení kandidátů. Nad touto mezí je kandidát považován za vlastní jméno osoby.

Výstupem celého bloku je seznam identifikovaných vlastních jmen osob.

3.2 Identifikace osob

Tento blok se zaměřuje především na identifikaci osob a sjednocení 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 řešen s pomocí doplňkové informace, kterou tvoří *téma*, pro které odhalujeme sociální síť. Předpokladem tohoto řešení je, že v dané tématické oblasti se vyskytuje pouze jedna osoba s jedinečnou kombinací jméno - příjmení.

3.3 Detekce vazeb mezi osobami

Detekce spojitostí mezi osobami je prováděna na základě výskytu jmen těchto osob společně v jednotlivých vstupních dokumentech (WWW stránkách). Použitý algoritmus vychází z postupu uvedeného v [15] s drobnými úpravami. Výpočty jsou vždy vztaženy k množině dokumentů S nalezené vyhledávačem k danému tématu.

Pro stanovení významnosti vazby mezi osobami t_i a t_j byla vypočtena hodnota $h_{ijS} = k\,(w_{iS} + w_{jS}) + (1\text{-}k)\,p_{ijS}$, kde h_{ijS} je již zmíněná významnost vazby mezi osobami t_i a t_j, w_{iS} je váha (význam) osoby t_i (analogicky t_j), p_{ijS} je vypočtená síla vazby daná společným výskytem jmen obou osob, vše vztaženo k množině dokumentů S. Volitelný koeficient k z intervalu <0,1> umožňuje zdůraznit složku vycházející z významnosti osob ($k \rightarrow 1$) nebo složku založenou na síle dané vazby ($k \rightarrow 0$).

Výsledkem této fáze je seznam dvojic osob, které se vyskytují společně, včetně ohodnocení jejich vazby hodnotou h_{ijS}. Aby tento seznam nebyl příliš rozsáhlý (což by snižovalo přehlednost výstupu), je pro zařazení vazby do výstupu potřeba, aby pro každou vazbu (hranu budoucího grafu tématické sociální sítě) bylo $h_{ijS} > m$, kde m je uživatelem definovaná mezní hodnota.

3.4 Vizualizace výstupů

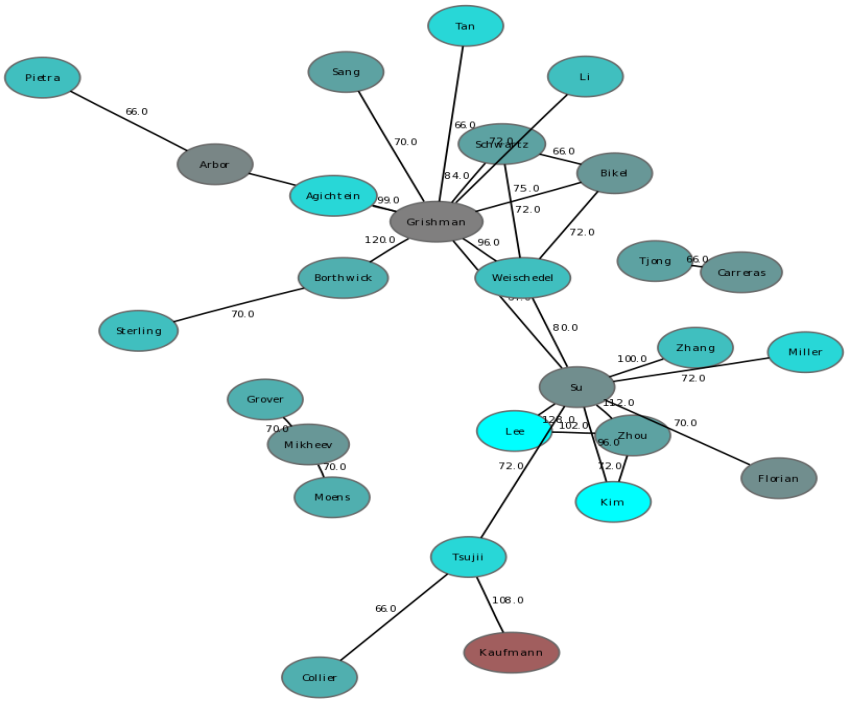
K vizualizaci celé tématické sítě je použita knihovna Graphviz [10]. Základním výstupem vizualizace je zobrazování osob a jejich vazeb z vybraných tématických oblastí. Rovněž je možné zobrazovat vazby vybrané osoby či seznamy osob souvisejících s daným tématem.

4 Architektura prototypu a dosažené výsledky

Výše uvedené postupy byly implementovány do prototypu webové aplikace. Ta umožňuje realizovat všechny činnosti nutné pro odhalení tématické sociální sítě od vyhledání relevantních WWW stránek pomocí vyhledávače Google a Google Scholar, přes detekci vlastních jmen, identifikaci osob a jejich vzájemných tématických vazeb až po vizualizaci výstupů.

Aby aplikace umožňovala rovněž automatizovaný režim provozu, kdy uživatel zadá seznam požadovaných témat k analýze a systém je zpracuje, je nutné stanovit několik uživatelsky definovaných koeficientů, aby systém dokázal co nejlépe odlišovat vlastní jména osob od jiných slov a slovních spojení. Proto je systém doplněn o moduly podporující stanovení těchto koeficientů na základě již analyzovaných témat a WWW stránek. Hlavním kritériem jejich stanovení je minimalizace chyby, které se bude systém dopouštět při klasifikaci kandidátů na vlastní jména osob.

S popsaným systémem byla provedena řada praktických experimentů. Ty odhalily jak pozitiva, tak i některá negativa celého řešení. Systém nyní obsahuje data z 1462 analyzovaných WWW stránek příslušných k 35 různým tématům. Počet klasifikovaných termů dosáhl 53778, z toho bylo 14041 pozitivně hodnocených jako vlastní jména osob. Jako příklad výstupu je na obr. 3 uveden graf identifikované tématické sítě pro téma „named entity recognition“ založené na analýze 81 webových stránek. Barva termů na obrazovce (odstín v tisku) odpovídá jejich váze, hrany jsou popsány významností dané vazby. Jak je patrné z obrázku, síť velmi dobře zachycuje vzájemné tématické propojení jednotlivých osob.



Obr. 3. Identifikovaná tématická síť pro téma „named entity recognition“

V průběhu experimentů byla také otestována metodika práce, která umožňovala dosahovat co nejlepší výsledky. Pro vybrané téma byl nejprve načten malý počet vstupních WWW stránek. Kandidáti na vlastní jména osob z těchto stránek byli klasifikováni ručně. Tak se aplikace seznámila s výrazy a slovními spojeními používanými v dané tématické oblasti a naučila se nepovažovat je za jména osob. Ve druhé fázi bylo již načítáno větší množství WWW stránek, přičemž systém byl již schopen provést detekci jmen s velmi dobrou přesností automaticky.

5 Shrnutí a další postup

Na základě provedených experimentů lze konstatovat, že pilotní aplikace vystavěná na zde popsaných metodách identifikace tématických sociálních sítí poskytuje zajímavé výsledky a vytváří vhodný základ pro další výzkum v této oblasti.

Další směry postupu se soustřeďují na oblasti, které by mohly zásadně ovlivnit šíři užití aplikace. Mezi ně patří například rozšíření vstupních importních filtrů. V současné době je primárním zdrojem dat vyhledávač Google, případně Google Scholar, a to především z důvodu zajištění obecného zdroje informací. Počítá se však s implementací dalších importních filtrů pro speciální data. Jako nejzajímavější pro vědeckou komunitu se zde jeví užití dat z citačních serverů.

V oblasti detekce vlastních jmen by bylo vhodné se zamyslet nad postupy stanovení hodnot uživatelsky definovaných koeficientů. Protože v některých případech jde v podstatě o optimalizační úlohu, je zvažováno použití genetických algoritmů.

Pro zobrazení výstupů by mohlo být užito 3D zobrazení pomocí jazyka VRML integrované v aplikaci, jinou možností by bylo využití některého existujícího vizualizačního nástroje s těmito možnostmi (např. Tulip).

6 Závěr

Popsané teoretické metody informačními technologiemi podporované identifikace tématických sítí jsou příspěvkem do aktuální oblasti detekce a užití sociálních sítí. Vytvořená pilotní aplikace je pak ukázkou konkrétního uplatnění teoretických metod 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.).

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. Computer and Information Science Papers CiteSeer Publications Research Index, In: *http://citeseer.ist.psu.edu/*, April 2008
5. DBIS-Homepage - DBL-Browser, In: *http://dbis.uni-trier.de/DBL-Browser/*, April 2008
6. DBLP Bibliography, In: *http://dblp.uni-trier.de/xml/*, October 2007
7. Frequently Occurring Names from the 1990 Census, In: *http://www.census.gov/genealogy/www/freqnames.html*, October 2007
8. Google, In: *http://www.google.cz/*, April 2008
9. Google Scholar, In: *http://scholar.google.cz/*, April 2008
10. Graphviz, In: *http://www.graphviz.org/*, October 2007.
11. 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.
12. ICU Project at the Data Privacy Laboratory, In: *http://privacy.cs.cmu.edu/dataprivacy/projects/icu/datainfo.html*, October 2007
13. ISI Highly Cited Researchers Version 1.5, In: *http://isihighlycited.com/*, April 2008
14. Jelínek, J.: Identifikace tématických sociálních sítí, konference Znalosti 2008, Bratislava, únor 2008, In: *Václav Snášel (Ed.): Znalosti 2008, pp. 90-100, FIIT STU Bratislava*, Ústav informatiky a softvérového inženierstva, 2008, ISBN: 978-80-227-2827-0
15. 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

16. 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
17. NetDraw Network Visualization, In: *http://www.analytictech.com/Netdraw/netdraw.htm*, April 2008
18. Stevenson, M.; Gaizauskas, R.: Using corpus-derived name lists for named entity recognition. In: *Proc. of ANLP*, Seattle, 2000.
19. Tulip Software home page, In: *http://tulip.labri.fr/*, April 2008
20. Web of Knowledge - ISI Web of Knowledge, In: *http://www.isiwebofknowledge.com/*, April 2008
21. WordNet, In: *http://www.cogsci.princeton.edu/~wn/*, October 2007.