Využitie sociálnych sietí pri vyhľadávaní v emailoch

Michal Laclavík, Ladislav Hluchý

Ústav informatiky, Slovenská akadémia vied,
Dúbravská cesta 9, 845 07 Bratislava
michal.laclavik@savba.sk

Abstrakt. Problém vyhľadávania v email archívoch pociťujú tak individuálni používatelia ako aj organizácie. V článku opisujeme prístup k vyhľadávaniu na základe extrakcie informácií, využitia sociálnych sietí ukrytých v emailoch ako aj použitia šírenia aktivácie za účelom hľadania relevantných relácií medzi objektmi obsiahnutými v emailoch.

Kľúčové slová: email, sociálne siete, vyhľadávanie informácií

1 Úvod

Email je stále najpoužívanejšou službou internetu [1]. Podľa najnovších štatistík [2] ľudia pracujúci s informáciami (tzv. information workers) odošlú a príjmu denne 110 emailových správ. Kým v minulosti počet prijatých emailov rástol, za posledných 5 rokov sú tieto štatistiky stabilné [3]. Email je stále populárnym a dôležitým nástrojom, pričom štatistiky z roku 2003 hovoria, že 80% užívateľov preferuje obchodnú komunikáciu cez email [4].

Informácie uložené v emailoch predstavujú hodnotu alebo príťaž v závislosti od toho ako dobré sú tieto informácie manažované. Email bol vymyslený za účelom asynchrónnej komunikácie, avšak v súčasnosti sa používa na veľa vecí na ktoré nebol vymyslený [5] [6], napríklad na upozornenia, informácie o transakciách, správu úloh, spoluprácu alebo archiváciu informácií. Týmto problémom emailovej komunikácie ako aj prehľadu súčasného stavu výskumu emailovej komunikácie sme sa venovali v našej predchádzajúcej práci [7].

V tomto článku sa zameriavame práve na úlohu archivovania informácií, na ktorú sa email používa. Pri archivovaní je potom dôležité vyhľadať informácie uschované v emailových archívoch organizácie alebo jednotlivca.

2 Sociálne siete a email

Emailové archívy obsahujú v sebe sociálnu sieť komunikujúcich adries, ľudí ako aj organizácií, či už ide o archívy firemné alebo osobné. Analýza a spracovanie emailových archívov nám umožňuje extrahovať a využiť túto sociálnu sieť, ktorá je napojená na ľudí, organizácie, geografické miesta (adresa), kontaktné informácie (telefón, email, adresa), čas alebo iné objekty. Takáto sociálna sieť ukrytá v emailových archívoch predstavuje hodnotný ale dosiaľ málo využitý zdroj pre organizácie alebo komunitu.

Sociálne siete a relevantné dáta v rámci sociálnych portálov ako Facebook sú vlastnené tretími stranami, pričom sociálne siete obsiahnuté v emailoch vlastní jednotlivec alebo organizácia, kde dáta sú väčšinou napojené na zaujímavé skryté fakty, ktoré je možné vyhľadať alebo odvodiť z týchto archívoch. V osobných

archívoch sa snaží sociálne siete využiť Xobni[2], ktorý na základe sociálnej siete manažuje kontakty a prílohy v emailoch. Avšak bolo by vhodné využiť sociálne siete na hľadanie vzťahov aj medzi inými objektmi obsiahnutými v archívoch organizácií alebo osobných email archívoch.

Sociálne siete v emailoch boli čiastočne analyzované. Napríklad v [8] sa autori zaoberali vzťahom CVS aktivity a aktivity na Apache Web Server mailingliste, kde zároveň riešili aj problém identifikácie email aliasov jednotlivých užívateľov. Extrakcia sociálnej siete a kontaktov z emailov a webu a ich kombinácia je diskutovaná v [9]. S extrakciou a transformáciou sociálnych sietí z emailov sme experimentovali aj my v našej práci [10]. Zaujímavým problémom je aj využitie jediného voľne dostupného firemného archívu emailov, Enron corpus, na analýzu sociálnych sietí [11] a vzťahov v nich obsiahnutých. V sociálne siete obsiahnutých v emailoch je možné objaviť úroveň interakcie (v čase, počet vymenených správ) a vzťahov (relácia k obsahu, sémantika). V našej práci sa snažíme využiť šírenie aktivácie na grafe multi-dimenzionálnej sociálnej siete podobne ako IBM Galaxy [12] kde bol predstavený koncept multi-dimensionálnej sociálnej siete na spracovanie textov.

V článku opisujeme ako sa dá tento prístup použiť na vyhľadávanie v emailoch.

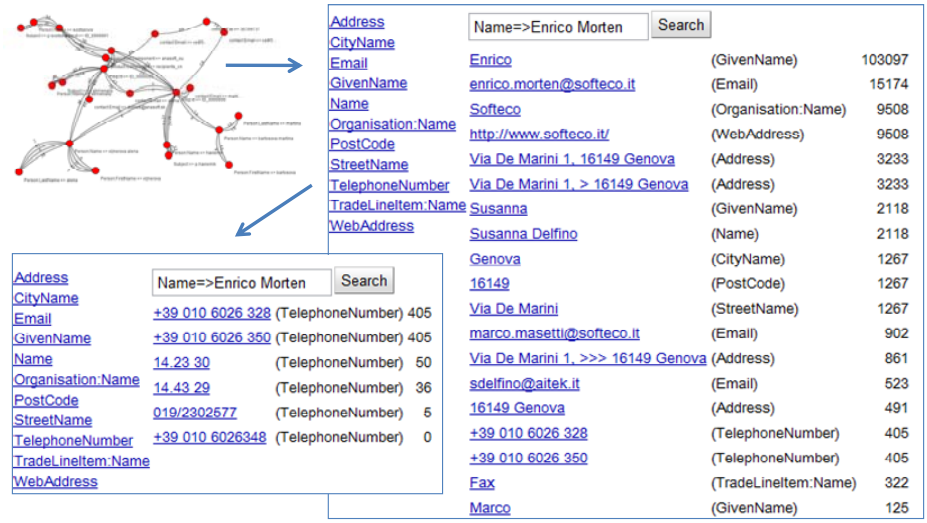
3 Email Social Network Search

V našej predchádzajúcej práci [13][14] sme vytvorili prototyp, ktorý extrahuje multi-dimenzionálne sociálne siete z emailov. Najskôr sa pomocou nástroja Ontea[3] nájdu objekty vo forme párov kľúč – hodnota, ktoré sa pre jednu správu organizujú do stromov. Takýto strom je potom zaradený do grafu (siete), kde pár kľúč - hodnota reprezentuje vrchol grafu. Takýto graf je zobrazený na obrázku č.1 vľavo hore. Na tieto grafy aplikujeme šírenie aktivácie tak ako bolo opísané v [14]. Výsledkom je že pre vrchol grafu dostaneme najrelevantnejšie vrcholy (objekty) obsiahnuté v email archívoch. Pričom tieto vrcholy nemusia byť priamo spojené s vrcholom z ktorého vychádza aktivácia.

V [13] a [14] sme vytvorili, opísali a vyhodnotili EmailSocialNetworkExtractor a algoritmus aktivácie ktorý je využitý v prototype Email Social Network Search. Úspešnosť (úplnosť aj pokrytie) algoritmu bola vyhodnotená medzi 60-77% [13] [14] a teda podobnú úspešnosť by mal dosahovať aj opísaný prototyp Email Social Network Serach, avšak pri zlepšení extrakcie by mohla byť aj vyššia.

Na obrázku 1, je zobrazené rozhranie vyhľadávania, kde vstupom je pár kľúč – hodnota reprezentujúca objekt (napr. človeka). Po vyhľadaní nám algoritmus šírenia aktivácie vráti relevantné objekty k tomuto objektu (človeku) ako napríklad telefónne čísla, organizácie, emailové adresy alebo mestá. (pozri obrázok 1 vpravo) Na obrázku 1 vľavo dole, môžeme vidieť predefinovanie tohto vyhľadávania, kedy nás zaujíma iba informácia určitého typu, teda obmedzíme vyhľadávanie kliknutím napríklad na kľúč (typ) telefónneho čísla, čo nám vráti relevantné telefónne čísla, k človeku.

[2] http://www.xobni.com/
[3] http://ontea.sourceforge.net/



Obr.1. Vyhľadávanie objektov v email archívoch. Vľavo hore: Multi-dimenzionálna sociálna sieť extrahovaná z emailov pomocou nástrojov Ontea a EmailSocialNetworkExtractor; Vpravo: relevantné výsledky rôznych typov vrátené k dopytu objektu človek. Vľavo dole: obmedzenie výsledkov na dopyt iba na telefónne čísla.

Súčasná verzia prototypu je vytvorená ako GWT[4] aplikácia. Vyhľadávanie musí začať vždy nejakým objektom reprezentovaným párom kľúč – hodnota, a teda funguje skôr ako vyhľadávanie pomocou navigácie a fazetového prehliadania, kde fazetou je kľúč (typy objektov). Zároveň je prototyp rozšírený o fulltextové vyhľadávanie, kde sú indexované objekty. Jednotlivé objekty je teda možné vyhľadať aj fulltextovo, pričom pri kliknutí na vrátené objekty pokračuje vyhľadávanie navigáciou.

Dôležitým plánovaným rozšírením je aj overenie správnosti informácie. Napríklad keď k človeku alebo firme vyhľadáme telefónne číslo, je dôležité aspoň čiastočne overiť či je to naozaj číslo človeka alebo firmy ktorému chceme zavolať. Môže sa totiž stať, že v archíve sa žiadne číslo na danú osobu nenachádza a systém nám odporučí číslo človeka, ktorá s danou osobou napríklad najviac komunikovala. Tento problém chceme riešiť rozšírením funkcionality o zobrazenie 3 najrelevantnejších okolí textu, alebo emailových správ, kde bolo telefónne číslo alebo iný objekt extrahovaný.

4 Záver

Pomocou vytvoreného prototypu je možné vyhľadávať súvislosti medzi objektmi, ktoré sú obsiahnuté v emailoch. Samozrejme tieto objekty musia byť najskôr objavené pomocou extrakcie informácií. Pravdepodobne by bolo vhodné doplniť riešenie o manuálne značkovanie zaujímavých objektov. Takto by si užívateľ mohol označkovať dôležité informácie ako heslá, linky, projekty a podobne, ktoré by potom boli dostupné pri vyhľadávaní. V budúcnosti plánujeme rozšírenie prototypu

[4] http://code.google.com/intl/sk-SK/webtoolkit/

o prístup k relevantným emailom kde sa informácie našli, o umožnenie vyhľadávania príloh ako aj vyhodnotenie prototypu na osobnej a firemnej emailovej korešpondencii. Súčasný prototyp extrakcie sociálnej siete tiež nie je dostatočne robustný a je potrebné riešiť aj perzistenciu extrahovanej siete. Zaujímavým sa javí aj rozšírenie grafu nie len na sieť extrahovanú z emailov, ale aj s iných dát, ako sú transakčné dáta, alebo dáta extrahované z dokumentov organizácie.

Poďakovanie: Táto práca vznikla s podporou projektov RECLER ITMS: 26240220029, SMART ITMS: 26240120005 a SMART II ITMS: 26240120029.

Literatúra

1. PewInternet report; Online Activities 2010; http://www.pewinternet.org/Static-Pages/Trend-Data/Online-Activites-Total.aspx, May 2010
2. The Radicati Group, Inc.: Email Statistics Report, 2010; Editor: Sara Radicati; http://www.radicati.com/wp/wp-content/uploads/2010/04/Email-Statistics-Report-2010-2014-Executive-Summary2.pdf
3. HP, The Radicati Group, Inc.: Taming the Growth of Email – An ROI Analysis (White Paper), http://www.radicati.com/wp/wp-content/uploads/2008/09/hp_whitepaper.pdf, 2005
4. META Group Inc.: 80% of Users Prefer E-Mail as Business Communication Tool http://www.mariosalexandrou.com/technology-trends/2003/80-percent-of-users-prefer-email.asp, 2003
5. S. Whittaker, C. Sidner: Email Overload: Exploring Personal Information Management of Email. In Proceedings of ACM CHI'96, 276-283, 1996
6. D. Fisher, A.J. Brush, E. Gleave & M.A. Smith: Revisiting Whittaker & Sidner's "email overload" ten years later. In CSCW2006, New York ACM Press, 2006
7. Michal Laclavík, Diana Maynard: Motivating intelligent email in business: an investigation into current trends for email processing and communication research; In E3C Workshop; IEEE CEC'10; DOI 10.1109/CEC.2009.47; pp. 476-482, 2009
8. Bird, C., Gourley, A., Devanbu, P., Gertz, M., Swaminathan, A., "Mining Email Social Networks", In: *MSR '06: Proceedings of the 2006 International Workshop on Mining Software Repositories*. ACM, New York (2006) 137–143.
9. Culotta, A., Bekkerman, R., McCallum, A.: "Extracting Social Networks and Contact Information from Email and the Web". In: *CEAS '04: Proceedings of the First Conference on Email and Anti-Spam*, 2004. http://www.ceas.cc/papers-2004/176.pdf
10. Michal Laclavík, Martin Šeleng, Ladislav Hluchý: Sociálne siete a E-mail; In Znalosti 2009: Vydavateľstvo STU, 2009. ISBN 978-80-227-3015-0, p. 313-316.
11. Diehl, C. P., Namata, G., Getoor, L., "Relationship Identification for Social Network Discovery" In: *The AAAI 2008 Workshop on Enhanced Messaging* (2008)
12. Judge, J., Sogrin, M., Troussov, A.: "Galaxy: IBM Ontological Network Miner" In: *Proceedings of the 1st Conference on Social Semantic Web*, Volume P-113 of Lecture Notes in In-formatics (LNI) series (ISSN 16175468, ISBN 9783-88579207-9). (2007)
13. Kvassay, M., Laclavík, M., Dlugolinský, Š.: "Reconstructing Social Networks from Emails". In: Pokorný, J., Snášel, V., Richta, K. (eds.): DATESO 2010: *Proceedings of the 10th annual workshop*, Prague (2010) 50-59. ISBN 978-80-7378-116-3
14. Michal Laclavík, Marcel Kvassay, Štefan Dlugolinský, Ladislav Hluchý: Use of Email Social Networks for Enterprise Benefit; In: International Workshop on Computational Social Networks (IWCSN 2010), *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2010