Reconstructing Social Networks from Emails

Marcel Kvassay, Michal Laclavík, and Štefan Dlugolinský

Institute of Informatics, Slovak Academy of Sciences
Dúbravská cesta 9, 845 07 Bratislava 45, Slovak Republic
marcel.kvassay@savba.sk

Abstract. The article provides a brief overview of Social Network Analysis (SNA) and its potential for exploiting the wealth of information buried in the email archives of business and private entities. Within the scope of the COMMIUS[1] project, we built a proof-of-concept prototype in Java, which used the spreading activation algorithm to reconstruct various aspects of multidimensional social network from emails. Two different variants of the spreading activation algorithm are discussed and compared.

1 Introduction

1.1 Social Network Analysis

History. The development of Social Network Analysis (SNA) is instructively mapped out by Freeman in [1] and, more briefly, by Fararo in [2]. Fig. 1 depicts the area at the intersection of sociology and psychology where SNA emerged as an alternative to traditional sociology. This area can be broadly classified as social psychology (SP), though it overlaps with other fields, such as anthropology or ethnology.

Psychology typically focuses on the individual. It studies mental functions – perception, cognition, attention, emotion, motivation, etc. – and their role in individual and social behavior. Sociology, in contrast, focuses on collective formations and analyzes human social activity starting from the micro level (agency and interaction) to the macro level (systems and social structures). In the area where sociology overlaps with psychology, the focus is on small groups. According to Wikipedia,[2] there are differences between social psychology as practised by psychologists (SP_p), and by sociologists (SP_s). Psychologists retain their individual focus and study how the thoughts, feelings, and behaviors of individuals are influenced by other members of the group. Sociologists focus more on the group itself and study group dynamics, crowd-phenomena, etc., often in the context of larger social structures (race, class, gender). Based on this distinction, the origins of SNA can be traced specifically to social psychology as practiced by sociologists and anthropologists (SP_s).

[1] http://www.commius.eu/

[2] http://en.wikipedia.org/wiki/Social_psychology

J. Pokorný, V. Snášel, K. Richta (Eds.): Dateso 2010, pp. 50–59, ISBN 978-80-7378-116-3.

Early forms of SNA, such as sociometry, appeared in the 1930s but for various reasons did not catch on. SNA was eventually accepted as a separate discipline in the 1970s, and has been on the rise ever since. With the spread of internet and cheap computing power, it penetrated mainstream sociology to such an extent that Wikipedia[3] now considers it "a key technique in modern sociology."

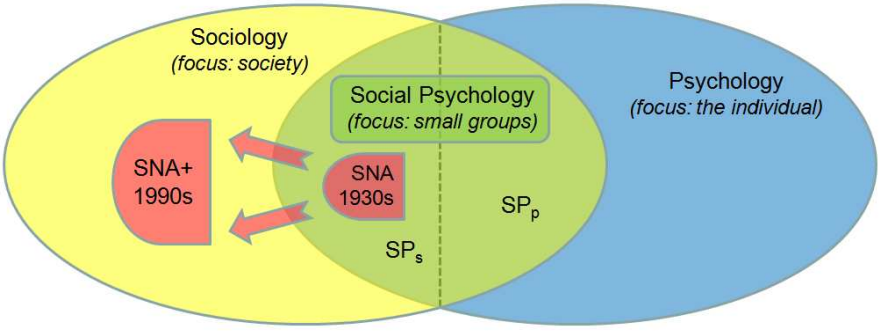

Fig. 1. SNA originated in the 1930s at the intersection of sociology and psychology, in the area shared by social psychology (SP), anthropology and ethnology. With the spread of internet and cheap computing power, it entered mainstream sociology (SNA+)

According to Freeman [1], SNA has four defining features:

1. Structural intuition (that patterning of social ties influences actors);
2. Systematic collection of empirical data;
3. Use of rigorous mathematical and computational models;
4. Graphic imagery.

Methods. To the use of statistics, which had a long tradition in sociology, SNA added new methods grounded in algebra and graph theory. SNA also enriched these disciplines with new concepts and techniques, such as block-modeling or the notions of bridge, centrality, structural balance, etc. More recently the focus has shifted to computational sociology and multi-agent simulations of social phenomena.

1.2 Social Networks in Emails

Email has become the most widespread Internet application. It is a tool supporting not only communication but also cooperation, task management, archiving, or information and knowledge management. Furthermore Email is a source of information on personal, enterprise or community network of an individual or an organization. Email communication analysis allows extraction of social networks with further connection to people, organizations, locations, topics or time.

Social Networks included in email archives are becoming increasingly valuable assets in organizations, enterprises and communities, though to date they have been little explored. While social networks in social network site such as Facebook are owned by

[3] http://en.wikipedia.org/wiki/Social_network_analysis

third parties, email social network data are owned by individual or organization including many useful connections hidden in emails. On personal archives, Xobni[4] exploits social networks to help the user manage contacts and attachments, but at the enterprise or community level, social networks can be exploited to improve email search, manage customers and suppliers, prioritise emails or improve inference mechanisms when connected with other detected semantic information from the email.

Social networks within email communication have been studied to some extent. For example, communication on the Apache Web Server mailing lists and its relation to CVS activity was studied in [6]. This work also introduces the problem of identifying email users' aliases. Extracting social networks and contact information from email and the Web and combining this information is discussed in [7]. Similarly, new email clients (e.g. Postbox) or plug-ins (Xobni) try to connect email social networks with web social networks like LinkedIn or Facebook. We have also performed some experiments on extraction of social networks from large email archives and network transformations using a semantic model [4]. Another research effort [8] exploits social networks to identify relations and tests proposed approaches on the Enron corpus.

To conclude, there is much research work done on social networks in the area of web social network applications, but email social networks are a bit different since in the email you can discover the level of interactions (number of messages exchanged, time, relation to content and possibly discovered semantics), and the influence of these differences on better information and knowledge management still needs to be explored. We would like to use similar approach as IBM Galaxy [10] in Nepomuk[5] project, where concept of multidimensional social network was introduced. In this paper we show initial results of exploiting email social network in order to support better understanding of email content as well as allowing applications such as partner or supplier search within organization or community.

1.3 COMMIUS[6] project

COMMIUS project is part of the 7th Framework Programme of the European Commission. Its acronym stands for "Community-based Interoperability Utility for Small and Medium Enterprises." The consortium comprises partners from Austria, Germany, Greece, Great Britain, Italy, Slovakia and Spain. Their objective is to provide SMEs (Small and Medium Enterprises) with "a zero, or very low-cost, entry into interoperability, based on non-proprietary protocols." To this end, a flexible architecture based on the open-source software was designed and implemented in Java.

The COMMIUS system is aimed primarily at companies that already conduct part of their business through email, i.e. send and receive orders, invoices, questionnaires, forms, etc. These documents are often manually retyped so as to enter them in the company's order-management or accounting system. Such companies would directly benefit from COMMIUS, since it is designed to automate these tasks. COMMIUS

[4] http://www.xobni.com/
[5] http://nepomuk.semanticdesktop.org/
[6] http://www.commius.eu/

scans the incoming emails, recognizes certain entities and documents (suppliers, customers, orders, invoices, telephone numbers, etc.), and proposes appropriate actions to the user. In case of an incoming order, for instance, the appropriate actions could be to check whether the requested items can actually be supplied, to send a confirmation to the customer, or to invoice and ship the order. These recommendations are presented to the user as HTML links that COMMIUS inserts into each incoming email. The user is then free to accept the recommendation (by clicking on the link) or proceed differently. A more detailed description of COMMIUS can be found in [3] and [4].

Social Networks in COMMIUS. The core of COMMIUS functionality deals with the business interoperability: detection of orders, invoices, payments, etc. in the incoming emails, and their semi-automated inclusion in the company's order-management and accounting system. Social network functionality is an add-on to this core. Its main purpose is to smoothen the process of COMMIUS adoption by new users, for example by pre-populating their product, customer and supplier databases based on the information extracted from their emails. On this basis, more advanced functions can be built, e.g. search for potential business partners.

Integration with the COMMIUS core. At the moment of installing COMMIUS, the user's email archives will be processed and the results stored in the form of multidimensional social network graph. After the installation, each incoming email will be added to this graph. Social network queries will search in the graph so that users get response in reasonable time. It should be noted that social network functions provide probabilistic results and would be offered to the user as recommendations only.

2 Implementation

We implemented our "social network extractor" in Java on top of the open-source graphical library JUNG.[7] The novelty of our approach is in the application of the spreading activation algorithm to two principal tasks:

1. reconstructing the social network from emails;
2. efficiently searching the social network.

The prototype implementation described below is a work in progress. So far only the initial version of the prototype has been tested on the first type of task; details are provided in section *3. Evaluation.*

2.1 Initial Version: Simple Cumulative Scorer

In the initial version, the information extraction module (IE) passes on a collection of strings (objects) matched by regular expressions. The strings acquire a "type" according to the regular expression that matched them. In this way, they are categorized as email addresses, personal names, names of organizations, telephone numbers, etc. The

[7] http://jung.sourceforge.net/

email message itself is represented by a “message ID” object to which the other objects are connected in a star-like fashion. If the same string is found in several messages, it is connected to all of them. If the same string is found multiple times in the same message, it has multiple links to that particular message ID, which is our way of recording the strength of the bond. The resulting network graph can be represented as a three-tier structure or tripartite graph (Fig. 2).

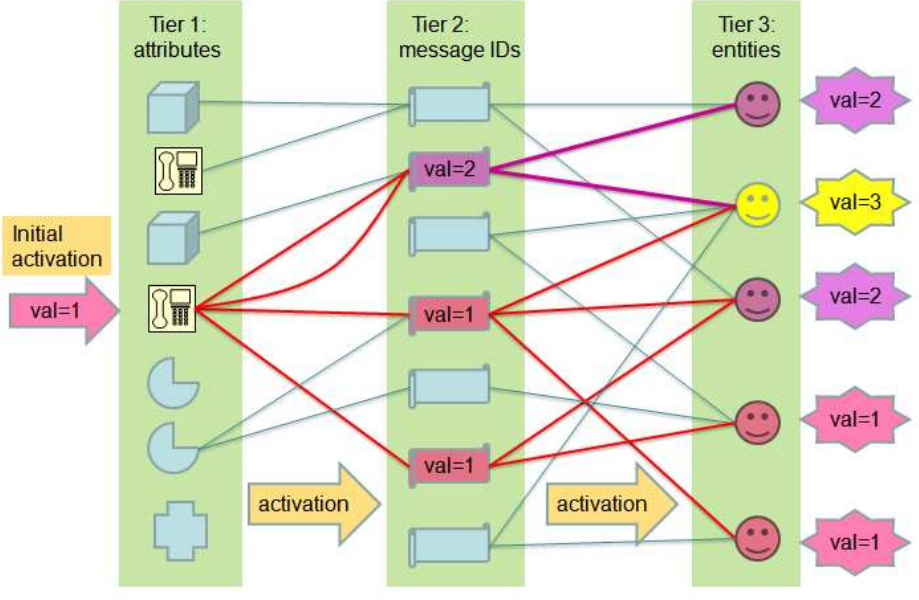

Fig. 2. Spread of activation as implemented in the initial version of the prototype. The activation starts from the left by assigning the initial activation value (val=1) to one attribute instance in Tier 1. In the next step, this initial value flows towards Tier 2 via the four red-colored links. There it accumulates in the three messages in which this attribute instance was found. One of them is connected via the double parallel link (which means it contained two occurrences of this attribute instance), so it accumulates a higher value (val=2) than the other two messages. In the final step, the activation values flow separately from each activated message (with val>0) towards Tier 3, where they further accumulate in the primary entities found in these messages. The primary entity with the highest accumulated value (val=3) is declared the “owner” of this attribute instance. The whole process is repeated for all the attribute instances in Tier 1

Our Simple Cumulative Scorer is inspired by the spreading activation algorithm in the sense that it separately “activates” each attribute instance with a uniform value of 1 and cumulatively spreads this activation (in the breadth-first manner) via the message IDs to the primary entities that will “own” the attribute. Each node can only fire once. It is an extremely simple implementation – we omitted such features of the standard spreading activation algorithm as attenuation, activation threshold or limit on the maximum activated value –nevertheless it allowed us to establish the utility of spreading activation in social networks.

In the three-tier structure depicted on Fig. 2, the middle tier (message IDs) links the attributes in Tier 1 to their "primary entities" in Tier 3. In general, the "primary entity" can be any of the objects identified by the Information Extractor (IE) if it can "own" (or be composed of) some other objects (attributes) likewise identified by the IE. In this sense, the "date" as a complex data type can be the primary entity with respect to year, month and day of which it is composed (provided the IE identifies these as separate objects), but it can itself be the attribute of a more complex data type, such as "event," "conference," etc. The number of such scenarios is limited only by the capabilities of the Information Extractor. In our case – since we are trying to reconstruct the social network – the primary entities are persons and organizations represented either by their email addresses or by their proper names. The Information Extractor collects both the email addresses and proper names from all the parts of the email message (the headers as well as the body). For our test task, we have chosen the telephone numbers as the sample attribute that we wish to assign correctly to persons and organizations.

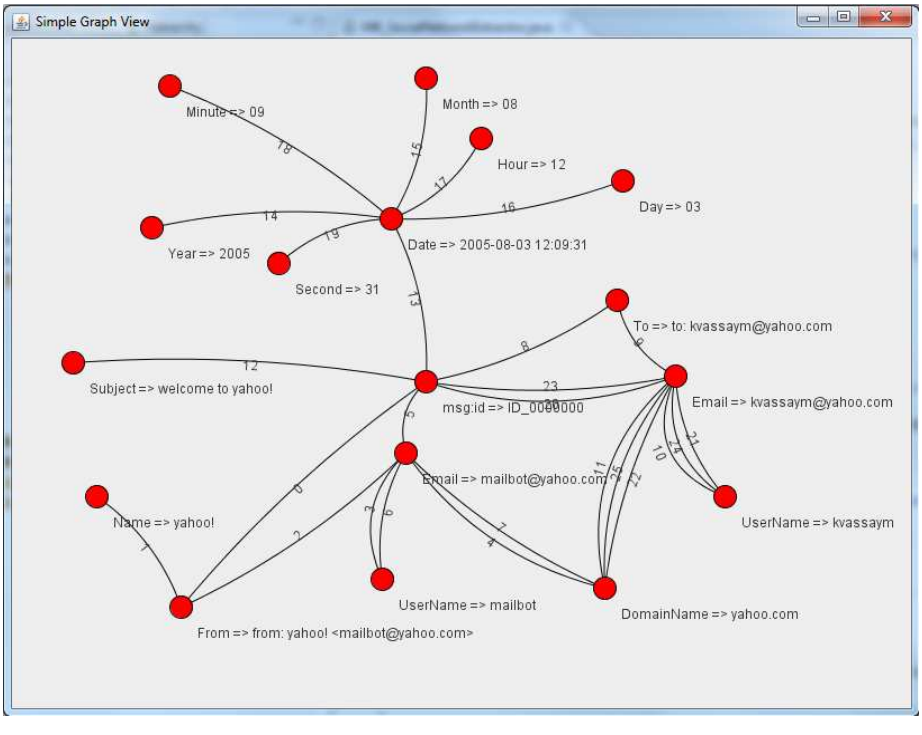

Fig. 3. Hierarchical graph of objects extracted by the enhanced information extractor (IE) from one email. Nodes representing sentences, paragraphs and blocks of the message are omitted. Graph was built from our personal email. Due to privacy reasons we cannot show the graph built from the enterprise emails of our Commius partners

2.2 Cumulative Edge Scorer with Attenuation

The improved version of the prototype relies on the enhanced output from the information extractor (IE). There are two major improvements:

1. IE produces a hierarchical tree of complex objects as shown on Fig. 3;
2. Objects are not linked directly to message IDs, but to the nodes representing the sentences, paragraphs and blocks of the message in which they were found. In this way the information about the physical proximity of the objects in the original email is preserved for further analysis as shown on Fig. 4.

In this richer and deeper tree-like structure, it makes sense to use a more sophisticated variant of spreading activation with attenuation and activation threshold. The structure can still be visualized as a multipartite graph, but the objects of each data type now require a separate partition (Fig. 4). This applies to candidate attributes as well as to primary entities. When partitioned in this way, the objects in each partition (i.e. the objects of the same data type) still have no connections among themselves, only to objects in other partitions, which is advantageous from the point of view of computational complexity.

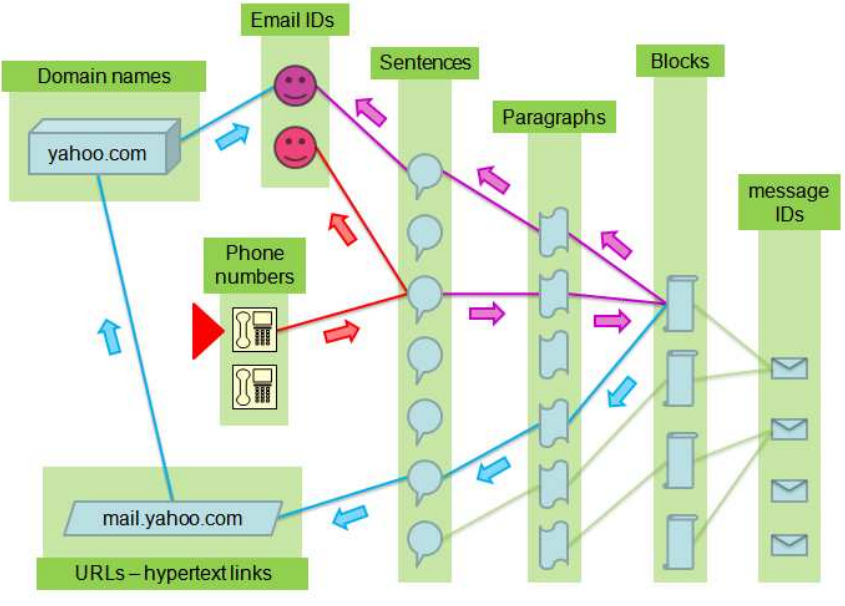

Fig. 4. A more complex variant of spreading activation in a multipartite graph. Activation again starts at one attribute instance (phone number) indicated by the red arrow (center left) and flows towards candidate primary entities (email IDs) by a variety of ways. Since the activation gets attenuated each time it passes through an edge, the shortest path (indicated in red) will carry over the greatest increment. But the longer paths (such as those indicated in purple and blue) can be so numerous that – depending on the value of attenuation and other parameters – their accumulated contribution ultimately prevails

Even in the enhanced version of our prototype, the spreading activation always starts from a single node. This allowed us to keep the simple and elegant breadth-first variant of the algorithm, in which each node can only fire once. Applications that start the initial activation from more than one node may need more sophisticated implementation of spreading activation.

3 Evaluation

We created a test email archive consisting of 28 representative sample messages supplied by our partners in COMMIUS. We then run the prototype with the task to assign telephone numbers to people and organizations (represented by their proper names) that were found in the emails.

Though our primary goal was to evaluate the spreading activation algorithm, we could not completely insulate it from the effects caused by the Information Extractor. These are seen primarily in the figures for the recall. Out of 17 unique relevant telephone numbers in the test emails, the IE identified 13. These 13 numbers were then passed on to the spreading activation algorithm. As can be seen from Table 1, the initial version of the spreading activation algorithm correctly assigned 8 of them, which resulted in the precision of 61.5% and the recall of 47%.

Table 1. Quantitative evaluation of the initial version of the spreading activation algorithm

Total relevant	Total found	Correctly assigned	Wrongly assigned	Recall [%]	Precision [%]
17	13	8	5	47	61.5

At present, we are still in the early stages of experimenting with the attenuated variant of the spreading activation. The first tentative results that we obtained are presented in Table 2.

Table 2. Quantitative evaluation of the attenuated variant of spreading activation

Total relevant	Total found	Correctly assigned	Wrongly assigned	Recall [%]	Precision [%]
17	13	10	3	58.8	76.9

As expected, the enhanced version of the prototype gave significantly better results (the precision of 76.9% and the recall of 58.8%), though we still need to investigate the remaining problems. Similarity in the tabular way of presenting the data actually masks deep differences between the two implementations.

In the initial version, the Information Extractor identified 32 candidate names of persons and organizations against which the phone numbers had to be matched. The candidate names were found purely by matching against regular expressions.

In the next version, it was decided to enhance the capabilities of the Information Extractor by adding gazetteers. Regular expressions were adapted so as to increase the recall and return a much larger number of candidate names, which would be subsequently filtered by the gazetteers. However, at the time of our experiment the gazetteers were not yet ready, so the enhanced version of the prototype had to match the phone numbers against a much larger set of 152 candidate names. That it still outperformed the initial version is therefore doubly significant and promising.

Moreover, the 3 phone numbers that were wrongly assigned had very low frequencies (one or two occurrences in the test corpus), so the basic assumptions of the algorithm were not met. Nevertheless, our initial experimentation with the prototypes was very useful and provided us with important hints for future work, which we discuss in the next section.

4 Conclusions and Future Work

This article is a report of a work in progress, and our experimentation with the prototypes still continues. The most obvious need is to further test the enhanced prototype on a larger set of representative emails. Here the main challenge is to get access to relevant and representative email sets. Though the email correspondence grows at an accelerating rate, we have learned that not all emails were created equal. They fall into distinct groups which differ significantly with respect to the ways and kinds of information that can be mined from them. Each application needs to be tested on the emails that are representative of the area in which it will be actually deployed.

The second lesson was that there were several alternative ways in which the Information Extractor could present the data extracted from the emails, and each had its pros and cons. There is a scope for deeper theoretical analysis here – either to find a general "canonical" structure suitable for most purposes or, alternatively, an easy way of transforming the data from one form to another depending on the task.

In general, graph and data transformations may be necessary in order to filter out the irrelevant information. Certain algorithms may require it for correct functioning; in others it will help to reduce the complexity of computation.

The "Social Network Extractor" component that we developed is able to process either mailboxes in mbox format or directories with email (.eml) messages, and thus extract multidimensional social network information contained in the email archive. In such a graph or network it is possible to see and exploit the links among objects such as people, time, email addresses, subjects, URLs, contact details or recipients.

The preliminary results of the extraction of social networks from email archives show that it is possible to deliver Xobni-like functionality in the enterprise or organizational context. Our approach is based on the concept of spreading activation similar to IBM Galaxy [10].

We have shown inferring relations between people and phone numbers on a small set of emails using a simple algorithm. The success rate (precision) of the experiment is 76.9%. In future, we would like to infer the relations such as those between custom-

ers and services, suppliers, products and transactions, organizations and people, people and address details, and others.

The extracted graph data from the email archives together with a well defined and tuned spreading activation algorithm can deliver the data needed for the adaptation of Commius or other enterprise systems. Such data can also be used to fill in the enterprise system database upon installation and thus help it to offer full functionality from the beginning. For example, we can populate a system database with a list of potential suppliers, organizations, contacts and their expertise.

Acknowledgements

This work is partially supported by projects Commius FP7-213876, APVV DO7RP-0005-08, AIIA APVV-0216-07, VEGA 2/0184/10 and VEGA 2/0211/09. We would also like to thank the anonymous reviewers, based on whose comments we have significantly reworked and enhanced the paper.

References

1. Freeman, L.: The Development of Social Network Analysis. Empirical Press, Vancouver (2006) (URL: http://aris.ss.uci.edu/~lin/book.pdf)
2. Fararo, T.J.: "Theoretical Sociology in the 20th Century." Journal of Social Structure 2. (2001) (URL: http://www.cmu.edu/joss/content/articles/volume2/Fararo.html)
3. Laclavík, M., Šeleng, M., Gatial, E., Hluchý, L.: Future email services and applications. In: CEUR-WS: Proceedings of the poster and demonstration paper track of the 1st Future Internet Symposium (FIS´08), Vol. 399. Telecon Res. Center, Vienna (2008) 33-35. ISSN 1613-0073. (URL: http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-399).
4. Laclavík, M., Šeleng, M., Ciglan, M., Hluchý, L.: Supporting Collaboration by Large Scale Email Analysis. In: Cracow´08 Grid Workshop: Proceedings. Academic Computer Centre CYFRONET AGH, Kraków (2009) 382-387. ISBN 978-83-61433-00-2
5. Balzert, S., Burkhart, T., Kalaboukas, K., Carpenter, M., Laclavik, M., Marin, C., Mehandjiev, N., Sonnhalter, K., Ziemann, J.: Appendix to D2.1.2: State of the Art in Interoperability Technology, Commius project deliverable (2009)
6. 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.
7. 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
8. Diehl, C. P., Namata, G., Getoor, L.: Relationship Identification for Social Network Discovery. In: The AAAI 2008 Workshop on Enhanced Messaging (2008)
10. 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 Informatics (LNI) series (ISSN 16175468, ISBN 9783-88579207-9). (2007)