Email Social Network Extraction and Search

Michal Laclavík, Štefan Dlugolinský, Marcel Kvassay, Ladislav Hluchý
Institute of Informatics, Slovak Academy of Sciences
michal.laclavik@savba.sk

Abstract

The article discusses our email search prototype, which exploits social networks hidden in email archives and the spread of activation algorithm. The prototype offers new way of searching email archives as knowledge repository. The prototype was partially evaluated on the Enron email corpus.

1. Introduction

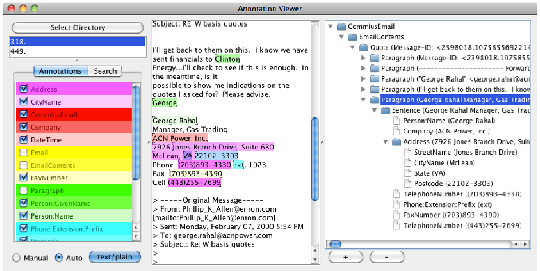
Email communication analysis allows the extraction of social networks with links to people, organizations, locations, topics or time. Social Networks included in email archives are becoming increasingly valuable assets in organizations, enterprises and communities, though to date they have been little explored.

Social networks in the area of email communication have been studied to some extent. Communication on the Apache Web Server mailing lists and its relation to CVS activity was studied in [1]. Extracting social networks and contact information from emails and the Web and combining this information is discussed in [2]. Another research effort [3] exploits social networks to identify relations, and tests the proposed approaches on the Enron corpus. We are using a similar approach to that of IBM Galaxy [4] in the Nepomuk[1] project, where the concept of multidimensional social network was introduced. Similar spread of activation inference was also tested on Wikipedia [7].

In the article we present a new approach for email search exploiting information extraction, graphs or networks and spread of activation. We build on our previous work [5] [6], where we presented an approach for the extraction, spreading activation and we also evaluated the relevance of the extraction and inference algorithm with satisfactory results. In this paper we discuss Email Social Network Search prototype, which uses the same algorithm as described in [5] and [6] but a new user interface for controlling the search has been built and we have also performed scalability evaluation on a part of the Enron email corpus[2].

2. Extraction of Social Network Graph

In order to provide the social network graph hidden in the email communication, the important task is to identify objects and object properties in the email and thus to formalize the message content and context. For object identification we use the information extraction (IE) techniques [6] based on regular expression patterns and gazetteers[3] as can be seen in Figure 1.


Annotation Viewer
Select Directory
318.
449.
Annotations
Search
☑ Address
☑ CityName
☑ CommiusEmail
☑ Company
☑ DateTime
☐ Email
☐ EmailContents
☑ FaxNumber
☐ Paragraph
☑ Person:GivenName
☑ Person:Name
☑ Phone:Extension:Prefix
○ Manual
● Auto
text/plain
CommiusEmail
EmailContents
Quote (Message-ID: <2398018.1075855692214
Paragraph (Message-ID: <2398018.1075855
Paragraph (---------------------- Forward
Paragraph ("George Rahal" <george.rahal@acn
Paragraph (I'll get back to them on this. I know
Paragraph (George Rahal Manager, Gas Tradir
Sentence (George Rahal Manager, Gas Trac
Person:Name (George Rahal)
Company (ACN Power, Inc.)
Address (7926 Jones Branch Drive, Suit
StreetName (Jones Branch Drive)
CityName (McLean)
State (VA)
Postcode (22102-3303)
TelephoneNumber ((703)893-4330)
Phone:Extension:Prefix (ext)
FaxNumber ((703)893-4390)
TelephoneNumber ((443)255-7699)
+
-

Subject: RE: W basis quotes

I'll get back to them on this. I know we have
sent financials to Clinton
Energy...I'll check to see if this is enough. In
the meantime, is it
possible to show me indications on the
quotes I asked for? Please advise.
George

George Rahal
Manager, Gas Trading
ACN Power, Inc.
7926 Jones Branch Drive, Suite 630
McLean, VA 22102-3303
Phone (703)893-4330 ext. 1023
Fax (703)893-4390
Cell (443)255-7699

> -----Original Message-----
> From: Phillip_K_Allen@enron.com
[mailto:Phillip_K_Allen@enron.com]
> Sent: Monday, February 07, 2000 5:54 PM
> To: george.rahal@acnpower.com
> Subject: Re: W basis quotes
>
>

Figure 1. User interface of Ontea [6] information extraction tool with highlighted extracted objects (middle) and tree structure (right) which is used to build social network graph (Figure 2).

Patterns and gazetteers extract key-value pairs (object type – object value) from email textual content as seen in the middle of Figure 1. If there is textual data present in binary form (e.g. PDF attachment) it is, if possible, converted to text before the information extraction begins. Extracted key-value pairs are then used to build the tree (see Figure 1 on the right side) and the graph of social network as can be seen in Figure 2. So the social network contains not only the communicating parties but also related entities, which can be explored.

[1] http://nepomuk.semanticdesktop.org/

[2] http://www.cs.cmu.edu/~enron/

[3] Gazetteer is simply the list of keywords (e.g. list of given names) representing an object type, which are matched against the text of emails.

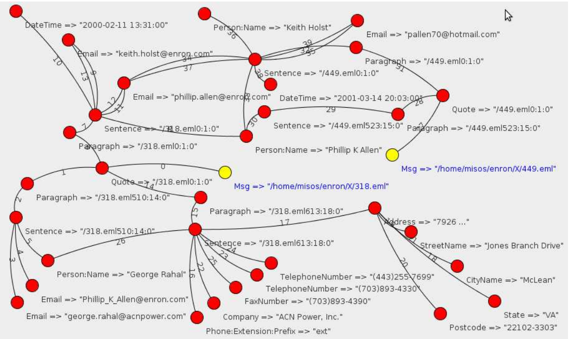
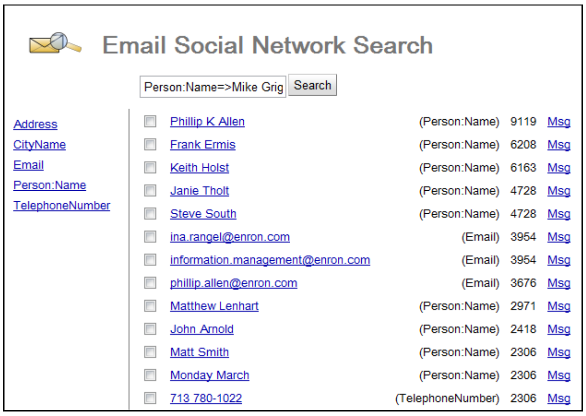

DateTime => "2000-02-11 13:31:00"
Person:Name => "Keith Holst"
Email => "pallen70@hotmail.com"
Email => "keith.holst@enron.com"
Paragraph => "/449.eml0:1:0"
Sentence => "/449.eml0:1:0"
Email => "phillip.allen@enron.com"
DateTime => "2001-03-14 20:03:00"
Quote => "/449.eml0:1:0"
Sentence => "/318.eml0:1:0"
Sentence => "/449.eml523:15:0"
Paragraph => "/449.eml523:15:0"
Paragraph => "/318.eml0:1:0"
Person:Name => "Phillip K Allen"
Msg => "/home/misos/enron/X/449.eml"
Quote => "/318.eml0:1:0"
Msg => "/home/misos/enron/X/318.eml"
Paragraph => "/318.eml510:14:0"
Paragraph => "/318.eml613:18:0"
Address => "7926 ..."
Sentence => "/318.eml510:14:0"
Sentence => "/318.eml613:18:0"
StreetName => "Jones Branch Drive"
Person:Name => "George Rahal"
TelephoneNumber => "(443)255-7699"
TelephoneNumber => "(703)893-4330"
CityName => "McLean"
Email => "Phillip_K_Allen@enron.com"
FaxNumber => "(703)893-4390"
Email => "george.rahal@acnpower.com"
Company => "ACN Power, Inc."
State => "VA"
Postcode => "22102-3303"
Phone:Extension:Prefix => "ext"

Figure 2. Social Network Graph built from 2 Enron emails. Please note that email or person name connects these 2 emails. Graph[4] also contains sentence and paragraph nodes (Figure 1, right). Here such nodes are hidden for readability.

3. Email Social Network Search Prototype

Email Social Network prototype exploits email analysis and extraction in the form of key-value pairs [6] (Figure 1, middle), semantic trees (Figure 1, right side) as well as their interconnection into a graph of social network [5] (Figure 2). It also exploits the spreading activation inference algorithm discussed in [5].

The idea is that we can infer relations of the activated object using spread of activation in social network graph. For example we can discover related objects for a person as seen in Figure 3.

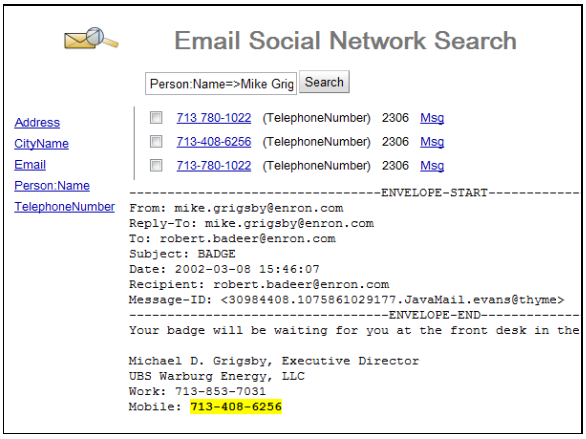


□	Phillip K Allen	(Person:Name)	9119	Msg
□	Frank Ermis	(Person:Name)	6208	Msg
□	Keith Holst	(Person:Name)	6163	Msg
□	Janie Tholt	(Person:Name)	4728	Msg
□	Steve South	(Person:Name)	4728	Msg
□	ina.rangel@enron.com	(Email)	3954	Msg
□	information.management@enron.com	(Email)	3954	Msg
□	phillip.allen@enron.com	(Email)	3676	Msg
□	Matthew Lenhart	(Person:Name)	2971	Msg
□	John Arnold	(Person:Name)	2418	Msg
□	Matt Smith	(Person:Name)	2306	Msg
□	Monday March	(Person:Name)	2306	Msg
□	713 780-1022	(TelephoneNumber)	2306	Msg

Figure 3. Email Social Network Search user interface. Relevant objects of various types returned for a person.

The Email Social Network Search Module enables the user to search real world objects mentioned in an email or email archive, and their relations. Exploiting the prototype, the user can discover personal email addresses, telephone numbers, company names, organizational names and so on. When the user searches for the object and accesses it, all the objects related to the searched object (i.e. personal names, phone numbers, organizations, addresses) are shown as results and the user can navigate deeper into the object graph by clicking on any object. For instance, the user can select a person's name to get the company where this person works, then select the company to get its phone number or address, and so on.

Figure 3 and 4 shows the user interface of the Email Social Network Search prototype. Its input is a key-value pair representing the business object (in this case, person) extracted from the email. When pressing the search button, the graph node in the search box is activated and the spreading activation algorithm is executed on the social network graph. The algorithm returns all the relevant objects (key-value pairs) for the searched object (Figure 3). For example, it returns organization names, email addresses, postal addresses, telephone numbers or personal names. Figure 4 shows the restriction of the search to return only the objects of a certain type (type is defined by the key of the key-value pair). In this case we wanted to get the phone number of the person, which was quite low in the main (mixed-type) list, so the search was further restricted to return only the relevant phone numbers for this person by clicking on the type *TelephoneNumber* in the left panel of GUI.



□	713 780-1022	(TelephoneNumber)	2306	Msg
□	713-408-6256	(TelephoneNumber)	2306	Msg
□	713-780-1022	(TelephoneNumber)	2306	Msg

```
--------------------------------ENVELOPE-START------------
From: mike.grigsby@enron.com
Reply-To: mike.grigsby@enron.com
To: robert.badeer@enron.com
Subject: BADGE
Date: 2002-03-08 15:46:07
Recipient: robert.badeer@enron.com
Message-ID: <30984408.1075861029177.JavaMail.evans@thyme>
---------------------------------ENVELOPE-END------------
Your badge will be waiting for you at the front desk in the

Michael D. Grigsby, Executive Director
UBS Warburg Energy, LLC
Work: 713-853-7031
Mobile: 713-408-6256
```

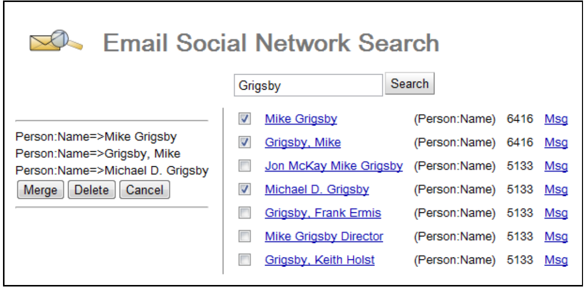
Figure 4. Email Social Network Search user interface. The same results as in Figure 3 restricted to phone numbers only. Phone numbers deemed relevant for the chosen person are displayed. Inferred objects (in this case phone numbers) can also be displayed and highlighted in their "most relevant" email message by clicking on "Msg" link.

In addition we can click on the "Msg" link next to the returned object in order to validate the search result. For example the prototype may return a telephone number of some other person (possibly somehow related) if the searched person does not have

[4] Values on edges are just edge IDs and their values are irrelevant for the algorithms.

any phone in the email archive. The requested object is then highlighted in the returned email message, which is the most relevant email inferred using the spread of activation, because the archive may contain hundreds of messages with the requested object.

The search algorithm can also be improved by allowing the users to delete the wrongly extracted objects or to connect various aliases of the same object (e.g. the same company or personal names spelled differently) as seen in Figure 5. Such a user feedback enables the search to learn and return better results in the future. For example, if we merge 3 selected person name aliases seen in Figure 5, there will be better results (e.g. phone number, address or organization) returned for any of the aliases.


Email Social Network Search

Grigsby

Person:Name=>Mike Grigsby
Person:Name=>Grigsby, Mike
Person:Name=>Michael D. Grigsby

☑	Mike Grigsby	(Person:Name)	6416	Msg
☑	Grigsby, Mike	(Person:Name)	6416	Msg
□	Jon McKay Mike Grigsby	(Person:Name)	5133	Msg
☑	Michael D. Grigsby	(Person:Name)	5133	Msg
□	Grigsby, Frank Ermis	(Person:Name)	5133	Msg
□	Mike Grigsby Director	(Person:Name)	5133	Msg
□	Grigsby, Keith Holst	(Person:Name)	5133	Msg

Figure 5. Prototype GUI with results of full-text search and several objects (aliases) selected for possible merge or delete. When these object aliases are merged, subsequent search returns better results for any of them.

In the future, we also plan to extend the prototype to let users annotate their own key-value pairs in the emails (using the Email Social Network Search GUI). In this way they can later easily search for the additional valuable information hidden in the emails, such as passwords, links or documents. In addition the annotated values will also be searched in other emails where they appear. An interesting feature would be to extend the search to include the attachments. We did not do it, since the version of the Enron corpus we have used did not include attachments, but in principle our prototype is now able to identify objects in attachments. The prototype and video explaining the search features is available at http://ikt.ui.sav.sk/esns/ and it works on several Enron mailboxes.

4. Evaluation

In [5] we have evaluated precision and recall of the information extraction and the spreading activation algorithms. The prototype and the algorithm were further refined with focus on higher precision of results. In this section we discuss mainly the scalability of the algorithm. Our hypothesis was that the performance (search time) should be stable even with large graphs, because we always activate only a small portion of the graph. This was found to be valid only to some extent.

Table 1. Email Social Network Search performance evaluation on 5 datasets.

Number of Mailboxes	1	5	7	10	15
Number of Emails	3 033	9 939	20 521	36 532	50 845
Number of Verticles	41812	159 776	369 932	608 146	835 025
Number of Edges	98566	380 254	971 929	1 796 403	2 514 031
Processing time (ms)	81 672	430 025	1 199 463	1 948 847	2 680 171
Processing time (minutes)	1	7	20	32	45
One Email processing time	27	43	58	53	53
Person:Name=>Mike Grigsby					
Search Response Time	**144**	**446**	**758**	**1 396**	**1 696**
Results	344	463	494	781	761
Fired	6 363	20 732	19 045	23 466	23 839
Visited	112 280	281 060	476 324	939 642	1 174 400
Visited Unique	18 382	53 772	82 219	145 192	178 829
Search Slowed down x Times	**1**	**3,1**	**5,3**	**9,7**	**11,8**
Fired x Times	**1**	**3,3**	**3,0**	**3,7**	**3,7**
Number of messages x Times	1	3,3	6,8	12,0	16,8
Number of verticles x Times	1	3,8	8,8	14,5	20,0
Number of edges x Times	1	3,9	9,9	18,2	25,5
TelephoneNumber=>713 780-1022					
Search Response Time	**5**	**8**	**8**	**12**	**13**
Results	4	4	4	4	4
Fired	116	150	157	181	183
Visited	6 318	8 776	9 550	13 424	14 710
Visited Unique	698	954	1 059	1 424	1 513
Search Slowed down x Times	**1**	**1,5**	**1,6**	**2,3**	**2,5**
Fired x Times	**1**	**1,3**	**1,4**	**1,6**	**1,6**
Number of messages x Times	1	3,3	6,8	12,0	16,8
Number of verticles x Times	1	3,8	8,8	14,5	20,0
Number of edges x Times	1	3,9	9,9	18,2	25,5
Address=>6201 Meadow Lake, Houston, TX 77057					
Search Response Time	**7**	**14**	**28**	**40**	**59**
Results	23	38	71	91	170
Fired	236	515	701	896	1 546
Visited	8 134	15 571	32 336	40 563	58 571
Visited Unique	1 097	1 952	6 526	8 029	11 295
Search Slowed down x Times	**1**	**2,1**	**4,3**	**6,0**	**8,9**
Fired x Times	**1**	**2,2**	**3,0**	**3,8**	**6,6**
Number of messages x Times	1	3,3	6,8	12,0	16,8
Number of verticles x Times	1	3,8	8,8	14,5	20,0
Number of edges x Times	1	3,9	9,9	18,2	25,5
Email=>ina.rangel@enron.com					
Search Response Time	**106**	**552**	**1 162**	**2 156**	**3 017**
Results	732	1 764	2 668	2 809	2 952
Fired	5 165	16 062	17 629	19 716	20 997
Visited	91 199	369 584	865 300	1 694 065	2 326 867
Visited Unique	13 355	54 987	81 757	134 876	168 955
Search Slowed down x Times	**1**	**5,2**	**11,0**	**20,3**	**28,5**
Fired x Times	**1**	**3,1**	**3,4**	**3,8**	**4,1**
Number of messages x Times	1	3,3	6,8	12,0	16,8
Number of verticles x Times	1	3,8	8,8	14,5	20,0
Number of edges x Times	1	3,9	9,9	18,2	25,5

One problem is that the social network graph has properties of small world networks. For example, in a similar work performed on Wikipedia graph [7], only two iterations of spreading activation could be performed because otherwise it would visit too many nodes. In [5] we have used 30 iterations, but on large graphs the impact on performance was too high. Now we set up the number of iterations to 4 experimentally. It seems to have no or minimal impact on the relevance of the returned results. The second problem is the implementation of our algorithm, which even with 4 iterations seems to visit too many nodes without firing the value (row *Visited*) and it visits the same nodes several times (compare with row *Visited Unique*).

As can be seen in Table 1, we have tested the search response time on 5 datasets from the Enron corpus. Datasets contained from 3,000 up to 50,000 emails. The biggest dataset resulted in a graph of 800,000 nodes and 2.5 million edges.

Before processing, graphs were loaded in memory. We have used Jung[5] library for graph representation. We plan to test email search also with SGDB [8], which would allow us to work with the whole Enron corpus or even larger email archives and test different versions of the spreading activation algorithm.

We have performed the same searches on all the five datasets for 4 different objects: person, telephone number, address and email address as seen in Table 1. The response time was computed as the average from 3 searches. We expected that the *Search Response Time* would not increase much with the number of nodes in the graph, but this was valid only partially – for the phone number and the address, which were deeper in the graph and had fewer connecting edges. Nodes such as emails or person names appear in emails more often and thus already 4 iterations of spreading activation activate (and especially visit) too many nodes for them. We believe that the problem is also in the implementation of our algorithm. The results show that the counts of the *Fired* and *Visited Uniqe* nodes do not increase so much with the number of nodes and edges in the graph, but the number of *Visited* (processed) nodes increases quite heavily. We believe this can be overcome in a better implementation, which would lead to a significantly better performance. In addition, we probably need to change the number of iterations of spread of activation depending on the topology of the graph near the activated nodes, e.g. to allow less iterations for nodes with many edges.

5. Conclusion

In this paper we discussed new approach to email search exploiting Email Social Networks based on spread of activation and information extraction. We have tested and partially validated our prototype on the Enron email corpus. We believe that the developed search interface, which allows user interaction with social network graph, offers innovative ideas for searching email archives as knowledge repository.

The search response time increases with the number of emails and nodes in the graph in some cases. We have tested it with 5 datasets from 3,000 up to 50,000 emails, where on some objects (person, email) the response time was slow and not so satisfactory, but on other object types such as phone or address the response time kept very good. We have suggested possible improvements, which will be part of our future work. We believe we can improve the response time through fine-tuning the spread of activation algorithm. The prototype and video is available at the web for possible user testing and feedback.

We believe we have shown new innovative approach for email search, which can be connected with data coming from other sources such as transactional databases, web, LinkedData or Wikipedia and thus offer searchable knowledgebase for enterprise, community or personal needs.

Acknowledgments

This work is supported by the projects ITMS: 26240220029 and VEGA 2/0184/10.

6. References

[1] Bird, C., Gourley, A., Devanbu, P., Gertz, M., Swaminathan, A., “Mining Email Social Networks”, In: *MSR ’06: Proceedings of the 2006 Workshop on Mining Software Repositories*. ACM, New York (2006) 137–143.

[2] Culotta, A., Bekkerman, R., McCallum, A.: “Extracting Social Networks and Contact Information from Email and the Web”. In: *CEAS’04*. http://www.ceas.cc/papers-2004/176.pdf

[3] Diehl, C. P., Namata, G., Getoor, L., “Relationship Identification for Social Network Discovery” In: *The AAAI 2008 Workshop on Enhanced Messaging, AAAI Conference On Artificial Intelligence*, pp 546-552 (2008)

[4] 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)

[5] Laclavík M., Kvassay M., Dlugolinský Š., Hluchý L.: Use of Email Social Networks for Enterprise Benefit; In: IWCSN 2010, IEEE/WIC/ACM WI-IAT, 2010, pp 67-70, DOI 10.1109/WI-IAT.2010.126

[6] Laclavík M., Dlugolinský Š., Šeleng M., Kvassay M., Gatial E., Balogh Z., Hluchý L.: Email Analysis and Information Extraction for Enterprise Benefit; In Computing and Informatics, 2011, vol. 30, no. 1, p. 57-87.

[7] Ciglan M., Nørvåg K.: WikiPop - Personalized Event Detection System Based on Wikipedia Page View Statistics (demo paper), Proceedings of CIKM'2010, Toronto, Canada, October 2010.

[8] Ciglan M., Nørvåg K.: SGDB - Simple graph database optimized for activation spreading computation, Proceedings of GDM'2010 (in conjunction with DASFAA'2010)

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