Email Social Network Extraction and Search

Michal Laclavík, Štefan Dlugolinský, Marcel Kvassay,
Ladislav Hluchý

IKT Group - Institute of Informatics SAS

Dept. of Parallel and Distributed Computing

Research and Development Areas:

- Large-scale HPCN and Grid applications
- Intelligent and Knowledge oriented Technologies

Experience from European IST projects:

- **3 project in** FP5**: ANFAS, CrosGRID, Pellucid**
- **6 project in** FP6**: EGEE II, K-Wf Grid, DEGREE (coordinator), EGEE, int.eu.grid, MEDIGRID**
- **4 projects in FP7:**
  Commius, Admire, EGEE III, Secricom

Several National Projects (SPVV, VEGA, APVT)

IKT Group Focus:

- Information Processing
- Semantic Web
- Knowledge oriented Technologies
- Parallel and Distributed
  Information Processing

Solutions:

- Ontea: Pattern-based Semantic Annotation
- ACoMA: KM tool in Email
- EMBET: Recommendation System

URL: http://ikt.ui.sav.sk

Director & leader of PDC:
Dr. Dipl. Ing. Ladislav Hluchý

Outline

- Social Networks in Emails
- Ontea: Information Extraction
- Business objects in Email Communication
- Building of Email Social Network
- Spread of Activation
- Relation Identification
- Email Social Network Search
- User Interaction with Data
- Evaluation

Motivation and Approach

Motivation

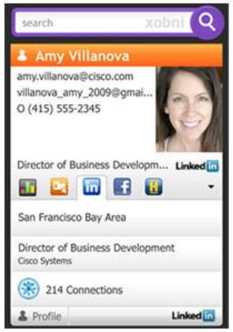
- To exploit information and knowledge included in email communication

Approach

- Social Network Extraction
- Entities extraction like People, Organizations, Locations, Contact data
- Forming semantic trees and graphs
- User interaction with graph data

Email Social Networks

- Email Social Networks are less explored
  - Several scientific publications:
    Apache mailing list, Enron, …
  - Commercial: Xobni (contacts and attachments)



search xobni

Amy Villanova

amy.villanova@cisco.com
villanova_amy_2009@gmai...
O (415) 555-2345

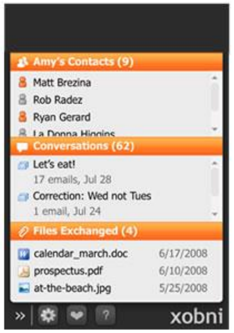
Director of Business Developm... LinkedIn

San Francisco Bay Area

Director of Business Development
Cisco Systems

214 Connections

Profile LinkedIn



Amy's Contacts (9)

Matt Brezina
Rob Radez
Ryan Gerard
La Donna Higgins

Conversations (62)

Let's eat!
17 emails, Jul 28
Correction: Wed not Tues
1 email, Jul 24

Files Exchanged (4)

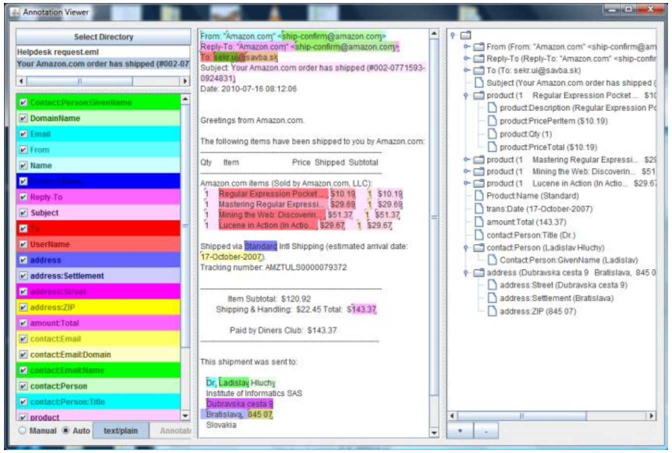
calendar_march.doc	6/17/2008
prospectus.pdf	6/10/2008
at-the-beach.jpg	5/25/2008

» xobni

- Benefit
  - Web Social Network Sites: owned by third parties
  - Email SN: owned by organization, individual or community
  - Additional level of interaction and context is present in emails
- Information and Knowledge
  - People, locations, contacts, product, services, attachments or links
  - Interactions
  - Time
  - Discovering relations can bring significant benefits
  - Spread of Activation – simple way to discover relations

Ontea: Information Extraction (Features)

- Regex patterns
- Visual Annotation Tool
- Integration with external tools
  - GATE, Stemers, Hadoop …
- Gazetteers
- IE System configuration
- Automatic loading of extractors
- Patterns
- Multilingual tests
  - Spanish
  - Slovak
  - English
  - Italian



From: "Amazon.com" <ship-confirm@amazon.com>
Reply-To: "Amazon.com" <ship-confirm@amazon.com>
To: sekr.ui@savba.sk
Subject: Your Amazon.com order has shipped (#002-0771593-0924831)
Date: 2010-07-16 08:12:06

Greetings from Amazon.com.

The following items have been shipped to you by Amazon.com:

```
Qty   Item                      Price   Shipped  Subtotal

Amazon.com items (Sold by Amazon.com, LLC):
 1    Regular Expression Pocket ...  $10.19   1  $10.19
 1    Mastering Regular Expressi...  $29.69   1  $29.69
 1    Mining the Web: Discoverin...  $51.37   1  $51.37
 1    Lucene in Action (In Actio...  $29.67   1  $29.67
```

Shipped via Standard Intl Shipping (estimated arrival date: 17-October-2007).
Tracking number: AMZTULS0000079372

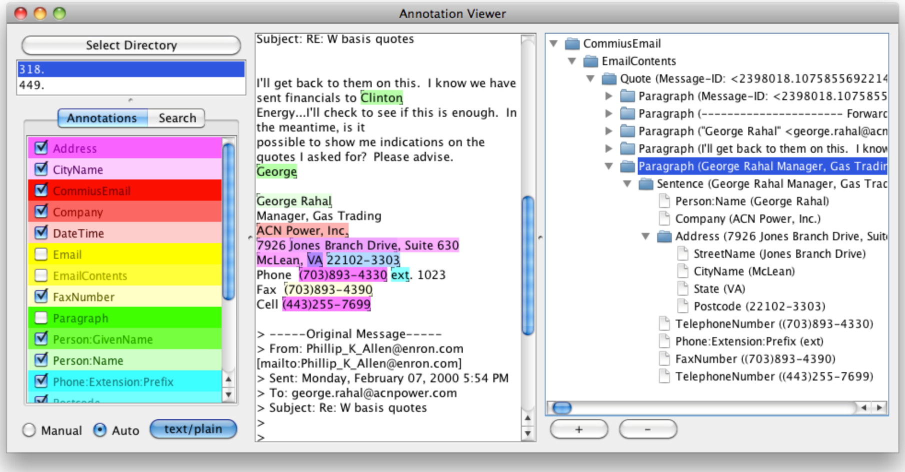
Item Subtotal: $120.92
Shipping & Handling: $22.45 Total: $143.37

Paid by Diners Club: $143.37

This shipment was sent to:

Dr. Ladislav Hluchy
Institute of Informatics SAS
Dubravska cesta 9
Bratislava, 845 07
Slovakia

From (From: "Amazon.com" <ship-confirm@am
Reply-To (Reply-To: "Amazon.com" <ship-confir
To (To: sekr.ui@savba.sk)
Subject (Your Amazon.com order has shipped (
product (1 Regular Expression Pocket ... $10
  product:Description (Regular Expression Pc
  product:PricePerItem ($10.19)
  product:Qty (1)
  product:PriceTotal ($10.19)
product (1 Mastering Regular Expressi... $29
product (1 Mining the Web: Discoverin... $51
product (1 Lucene in Action (In Actio... $29.67
ProductName (Standard)
trans:Date (17-October-2007)
amount:Total (143.37)
contact:Person:Title (Dr.)
contact:Person (Ladislav Hluchy)
  Contact:Person:GivenName (Ladislav)
address (Dubravska cesta 9 Bratislava, 845 0
  address:Street (Dubravska cesta 9)
  address:Settlement (Bratislava)
  address:ZIP (845 07)



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

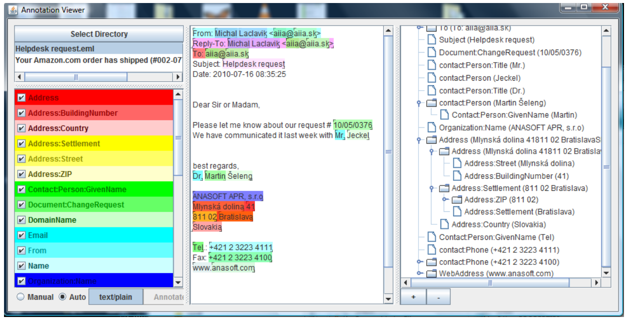
George Rahal
Manager, Gas Trading
ACN Power, Inc.
7926 Jones Branch Drive, Suite 630
McLean, VA 22102-3303
Phone (703)893-4330 ext. 1023
Fax (703)893-4390
Cell (443)255-7699

> -----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
>
>

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 Tradin
        Sentence (George Rahal Manager, Gas Trad
          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)

Business objects in Emails

- Study on 6 organizations show:
  - **Objects can be identified by patterns and gazeteers**
  - **It is possible to define set of common objects**
- Objects identified:
  - **Organization:**
    - org:Name, org:RegNo, org:TaxNo
  - **Person:**
    - person:Name, person:Function
  - **Contact:**
    - contact:Phone, contact:Email, contact:Webpage
  - **Address:**
    - address:ZIP, address:Street, address:Settlement
  - **Product:**
    - product:Name, product:Module, product:Component, product:BOID
  - **Document:**
    - doc:Invoice, doc:Order, doc:Contract, doc:ChangeRequest
  - **Inventory:**
    - inventory:ResID, inventory:ResType
  - **Other business object**
    - ID: BOID



Helpdesk request.eml
Your Amazon.com order has shipped (#002-07

☑ Address
☑ Address:BuildingNumber
☑ Address:Country
☑ Address:Settlement
☑ Address:Street
☑ Address:ZIP
☑ Contact:Person:GivenName
☑ Document:ChangeRequest
☑ DomainName
☑ Email
☑ From
☑ Name
☑ Organization:Name

From: Michal Laclavik <aiia@aiia.sk>
Reply-To: Michal Laclavik <aiia@aiia.sk>
To: aiia@aiia.sk
Subject: Helpdesk request
Date: 2010-07-16 08:35:25

Dear Sir or Madam,

Please let me know about our request # 10/05/0376
We have communicated it last week with Mr. Jeckel

best regards,
Dr. Martin Šeleng

ANASOFT APR, s.r.o
Mlynská dolina 41
811 02 Bratislava
Slovakia

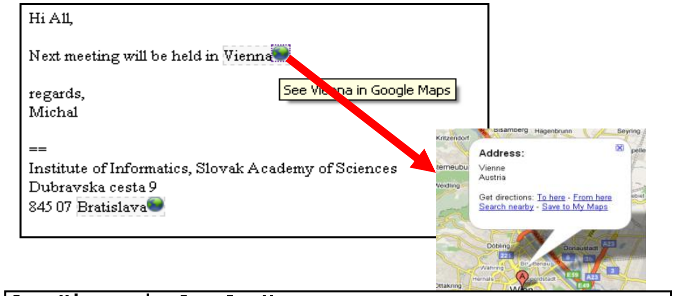
Tel.: +421 2 3223 4111
Fax: +421 2 3223 4100
www.anasoft.com

To (To: aiia@aiia.sk)
Subject (Helpdesk request)
Document:ChangeRequest (10/05/0376)
contact:Person:Title (Mr.)
contact:Person (Jeckel)
contact:Person:Title (Dr.)
contact:Person (Martin Šeleng)
  Contact:Person:GivenName (Martin)
Organization:Name (ANASOFT APR, s.r.o)
Address (Mlynská dolina 41811 02 BratislavaS
  Address (Mlynská dolina 41811 02 Bratisla
    Address:Street (Mlynská dolina)
    Address:BuildingNumber (41)
  Address:Settlement (811 02 Bratislava)
    Address:ZIP (811 02)
    Address:Settlement (Bratislava)
  Address:Country (Slovakia)
Contact:Person:GivenName (Tel)
contact:Phone (+421 2 3223 4111)
contact:Phone (+421 2 3223 4100)
WebAddress (www.anasoft.com)

Acoma architecture : Message Post Processing

Acoma is not part of Paper but related to NextMail

- **Useful hints with links are included in enriched email**
- **Links lead to internal or external systems (Internet, Intranet)**



Hi All,

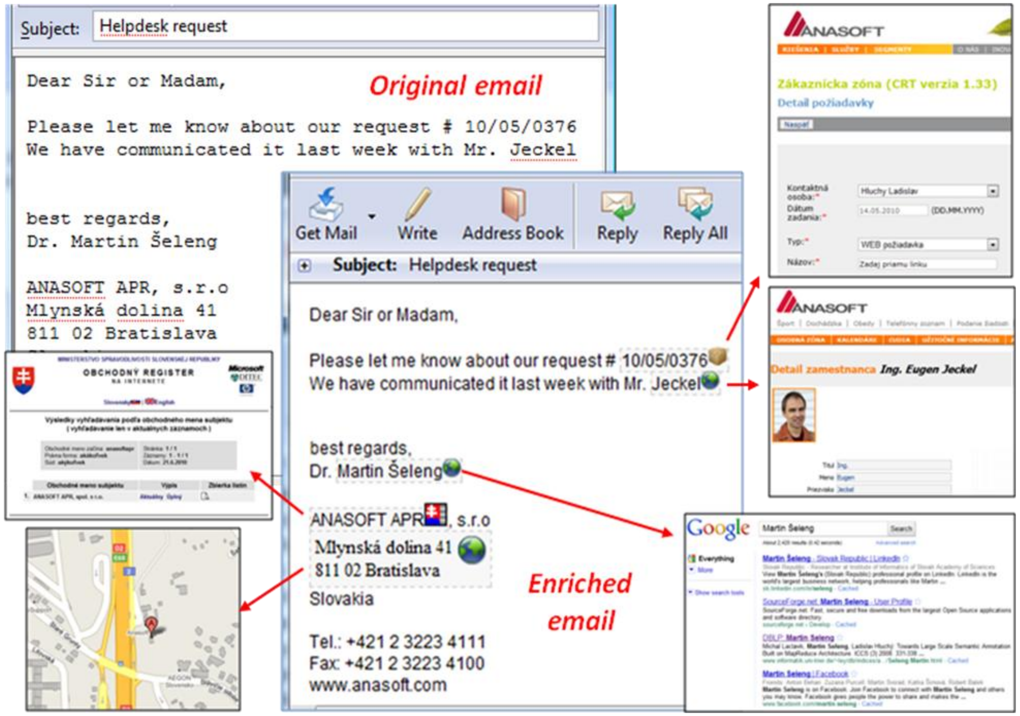
Next meeting will be held in Vienna.

regards,
Michal

==
Institute of Informatics, Slovak Academy of Sciences
Dubravska cesta 9
845 07 Bratislava

See Vienna in Google Maps
Address:
Vienne
Austria
Get directions: To here - From here
Search nearby - Save to My Maps

```
See Vienna in Google Maps
http://maps.google.com/maps?f=q&hl=en&geocode=&q=Vienna
Link,38,44
```



Subject: Helpdesk request

Original email

```
Dear Sir or Madam,

Please let me know about our request # 10/05/0376
We have communicated it last week with Mr. Jeckel

best regards,
Dr. Martin Šeleng

ANASOFT APR, s.r.o
Mlynská dolina 41
811 02 Bratislava
```

Subject: Helpdesk request

Dear Sir or Madam,

Please let me know about our request # 10/05/0376
We have communicated it last week with Mr. Jeckel

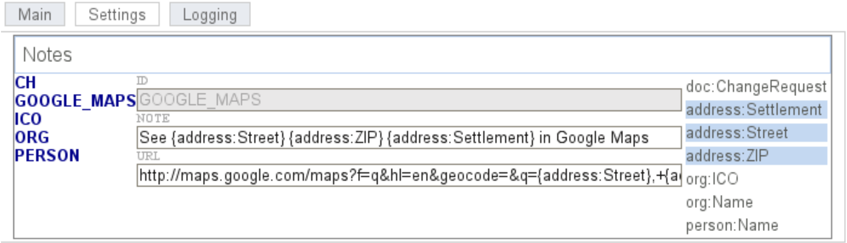
best regards,
Dr. Martin Šeleng

ANASOFT APR, s.r.o
Mlynská dolina 41
811 02 Bratislava
Slovakia

Tel.: +421 2 3223 4111
Fax: +421 2 3223 4100
www.anasoft.com

Enriched email

OBCHODNÝ REGISTER
ANASOFT
Zákaznicka zóna (CRT verzia 1.33)
Detail požiadavky
Kontaktná osoba: Hluchy Ladislav
Dátum zadania: 14.05.2010 (DD.MM.YYYY)
Typ: WEB požiadavka
Názov: Zadaj priamu linku
ANASOFT
Detail zamestnanca Ing. Eugen Jeckel
Titul Ing.
Meno Eugen
Priezvisko Jeckel
Google
Martin Šeleng
Search

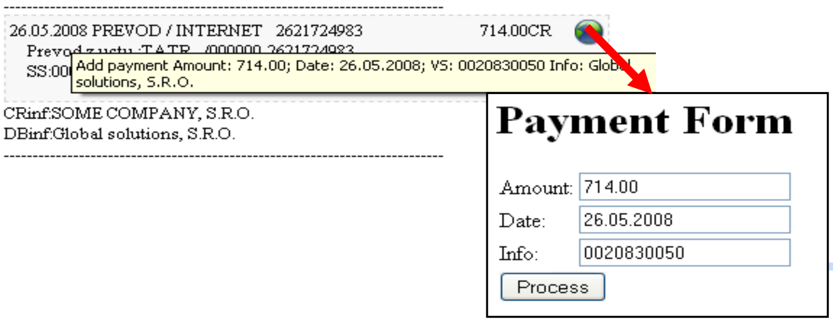


Notes

CH
GOOGLE_MAPS
ICO
ORG
PERSON

ID
GOOGLE_MAPS
NOTE
See {address:Street} {address:ZIP} {address:Settlement} in Google Maps
URL
http://maps.google.com/maps?f=q&hl=en&geocode=&q={address:Street},+{a

doc:ChangeRequest
address:Settlement
address:Street
address:ZIP
org:ICO
org:Name
person:Name



26.05.2008 PREVOD / INTERNET 2621724983 714.00CR
Prevod z uctu :TATR /000000 2621724983
SS:00

Add payment Amount: 714.00; Date: 26.05.2008; VS: 0020830050 Info: Global solutions, S.R.O.

CRinf:SOME COMPANY, S.R.O.
DBinf:Global solutions, S.R.O.

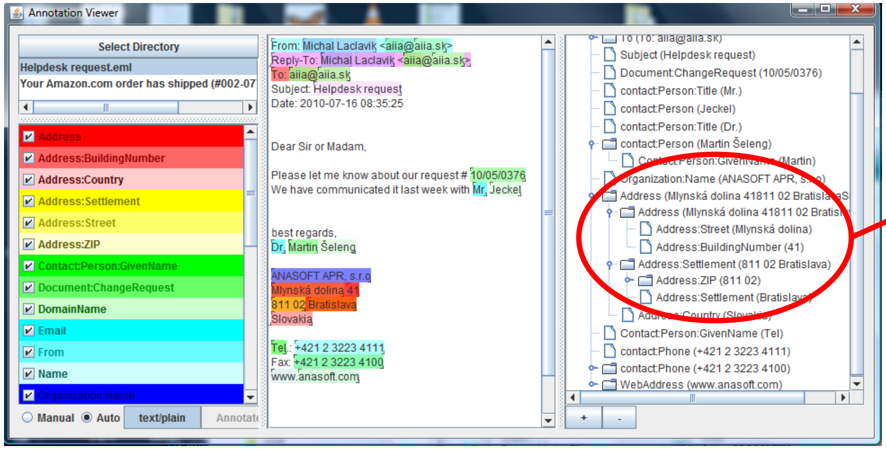
Payment Form

Amount: 714.00
Date: 26.05.2008
Info: 0020830050

Process

Acoma: Hint Recommendation

Acoma is not part of Paper but related to NextMail



Helpdesk request.eml
Your Amazon.com order has shipped (#002-07

- ☑ Address
- ☑ Address:BuildingNumber
- ☑ Address:Country
- ☑ Address:Settlement
- ☑ Address:Street
- ☑ Address:ZIP
- ☑ Contact:Person:GivenName
- ☑ Document:ChangeRequest
- ☑ DomainName
- ☑ Email
- ☑ From
- ☑ Name
- ☑ Organization:Name

○ Manual ● Auto

From: Michal Laclavik <aiia@aiia.sk>
Reply-To: Michal Laclavik <aiia@aiia.sk>
To: aiia@aiia.sk
Subject: Helpdesk request
Date: 2010-07-16 08:35:25

Dear Sir or Madam,

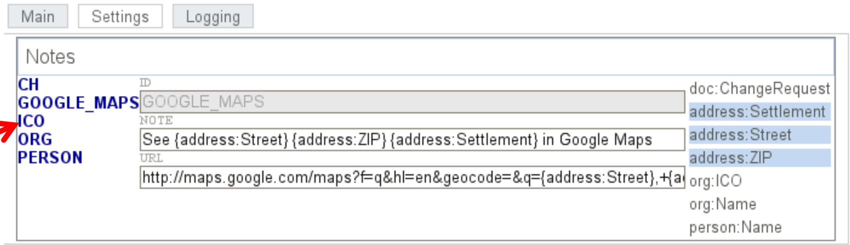
Please let me know about our request # 10/05/0376
We have communicated it last week with Mr. Jeckel

best regards,
Dr. Martin Šeleng

ANASOFT APR, s.r.o
Mlynská dolina 41
811 02 Bratislava
Slovakia

Tel.: +421 2 3223 4111
Fax: +421 2 3223 4100
www.anasoft.com

```
To (To: aiia@aiia.sk)
Subject (Helpdesk request)
Document:ChangeRequest (10/05/0376)
contact:Person:Title (Mr.)
contact:Person (Jeckel)
contact:Person:Title (Dr.)
contact:Person (Martin Šeleng)
    Contact:Person:GivenName (Martin)
Organization:Name (ANASOFT APR, s.r.o)
Address (Mlynská dolina 41811 02 BratislavaS
    Address (Mlynská dolina 41811 02 Bratisl
        Address:Street (Mlynská dolina)
        Address:BuildingNumber (41)
    Address:Settlement (811 02 Bratislava)
        Address:ZIP (811 02)
        Address:Settlement (Bratislava)
    Address:Country (Slovakia)
Contact:Person:GivenName (Tel)
contact:Phone (+421 2 3223 4111)
contact:Phone (+421 2 3223 4100)
WebAddress (www.anasoft.com)
```



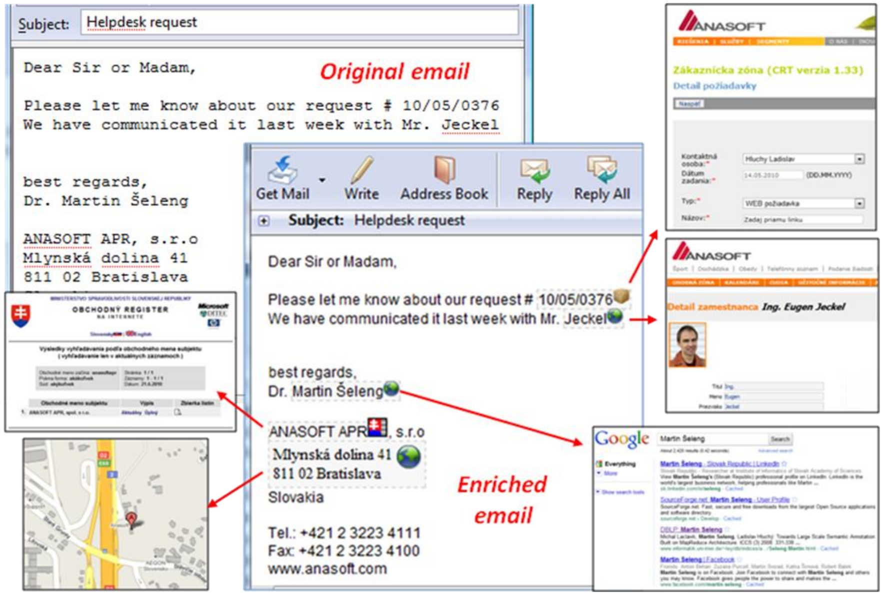
CH
GOOGLE_MAPS
ICO
ORG
PERSON

ID: GOOGLE_MAPS

NOTE: See {address:Street} {address:ZIP} {address:Settlement} in Google Maps

URL: http://maps.google.com/maps?f=q&hl=en&geocode=&q={address:Street},+{a

doc:ChangeRequest
address:Settlement
address:Street
address:ZIP
org:ICO
org:Name
person:Name



Subject: Helpdesk request

```
Dear Sir or Madam,

Please let me know about our request # 10/05/0376
We have communicated it last week with Mr. Jeckel


best regards,
Dr. Martin Šeleng

ANASOFT APR, s.r.o
Mlynská dolina 41
811 02 Bratislava
```

Subject: Helpdesk request

Dear Sir or Madam,

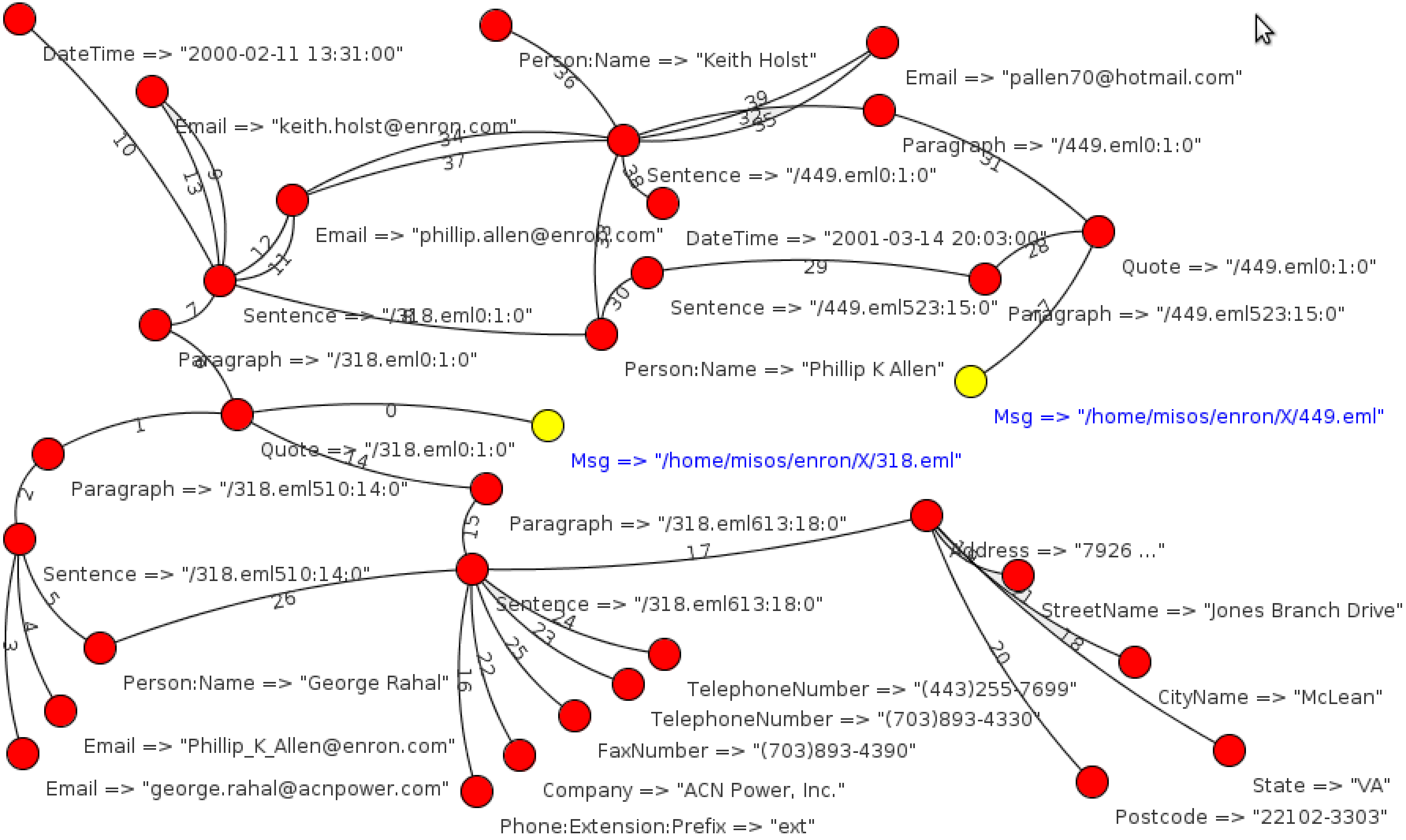
Please let me know about our request # 10/05/0376
We have communicated it last week with Mr. Jeckel

best regards,
Dr. Martin Šeleng

ANASOFT APR, s.r.o
Mlynská dolina 41
811 02 Bratislava
Slovakia

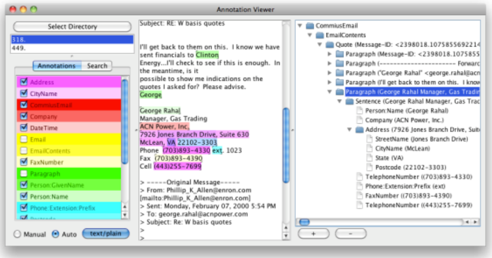
Tel.: +421 2 3223 4111
Fax: +421 2 3223 4100
www.anasoft.com

Email Social Graph/Network


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 => "/449.eml523:15:0"
Paragraph => "/449.eml523:15:0"
Sentence => "/318.eml0:1:0"
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"
CityName => "McLean"
TelephoneNumber => "(703)893-4330"
FaxNumber => "(703)893-4390"
Email => "Phillip_K_Allen@enron.com"
State => "VA"
Email => "george.rahal@acnpower.com"
Company => "ACN Power, Inc."
Postcode => "22102-3303"
Phone:Extension:Prefix => "ext"
10
13
9
36
39
32
35
34
37
31
38
12
11
28
29
30
7
0
1
14
2
15
17
26
5
4
3
16
22
25
23
24
20
18

Email Social Network Search: Features

- Social network of communicating people with relation to other entities
- Discovering relation in the graph/network using spread of activation
- Showing relations restricted to concrete type, e.g. telephone numbers related to a person
- User interaction with data (merging, deleting entities) with immediate impact on discovered relations
- Navigation over related entities
- Full-text search of the entities
- User interface for search



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

Subject: RE: W basis quotes

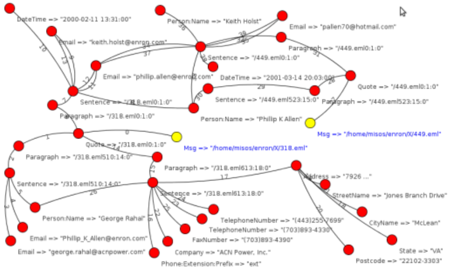
I'll get back to them on this. I know we have sent financials to Clinton Energy...I'll check to see if this is enough. In the meantime, is it possible to show me indications on the quotes I asked for? Please advise.
George

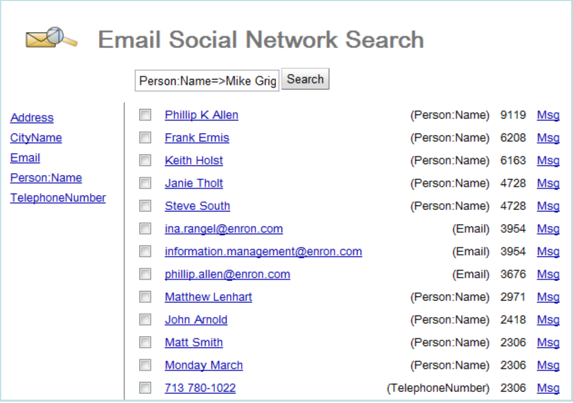
George Rahal
Manager, Gas Trading
ACN Power, Inc.
7926 Jones Branch Drive, Suite 630
McLean, VA 22102-3303
Phone (703)893-4330 ext. 1023
Fax (703)893-4390
Cell (443)255-7699

> -----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
>
>

```
▼ CommiusEmail
  ▼ EmailContents
    ▼ Quote (Message-ID: <2398018.1075855692214
      ▶ Paragraph (Message-ID: <2398018.1075855
      ▶ Paragraph (---------------------- Forwarc
      ▶ Paragraph ("George Rahal" <george.rahal@acn
      ▶ Paragraph (I'll get back to them on this.  I know
      ▼ Paragraph (George Rahal Manager, Gas Tradin
        ▼ 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)
```

+ −


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 => "Philip 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"
Phone:Extension:Prefix => "ext"
Postcode => "22102-3303"



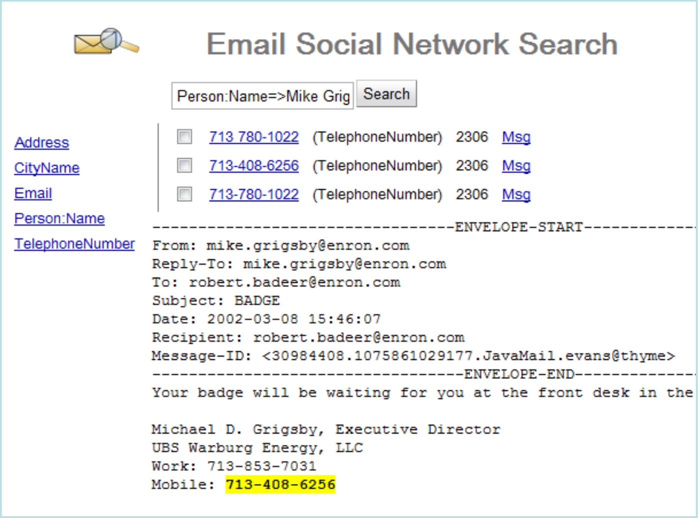
Email Social Network Search

Person:Name=>Mike Grig | Search

Address
CityName
Email
Person:Name
TelephoneNumber

☐	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

GUI Features

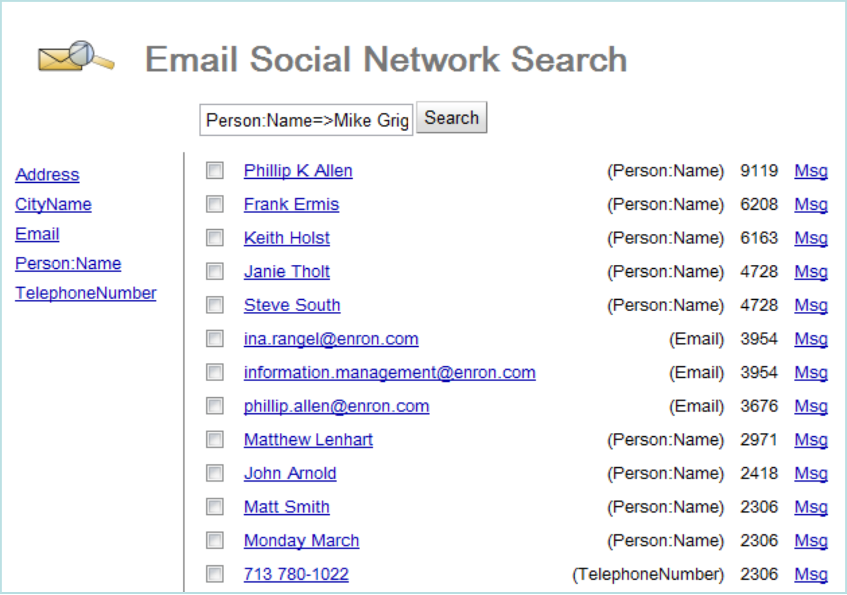


Person:Name=>Mike Grig

Address
CityName
Email
Person:Name
TelephoneNumber

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

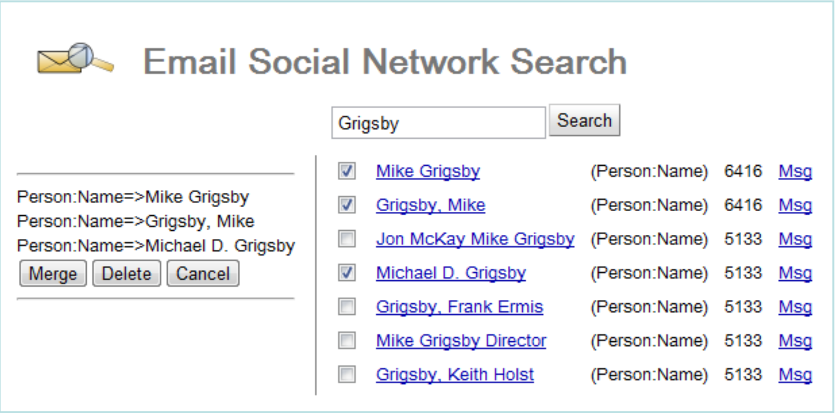
```
--------------------------------ENVELOPE-START------------
From: mike.grigsby@enron.com
Reply-To: mike.grigsby@enron.com
To: robert.badeer@enron.com
Subject: BADGE
Date: 2002-03-08 15:46:07
Recipient: robert.badeer@enron.com
Message-ID: <30984408.1075861029177.JavaMail.evans@thyme>
---------------------------------ENVELOPE-END-------------
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
```



Person:Name=>Mike Grig

Address
CityName
Email
Person:Name
TelephoneNumber

□	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

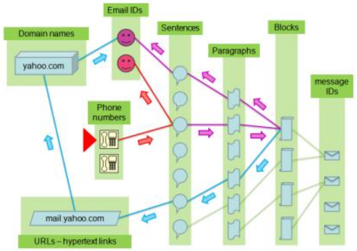


Grigsby

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

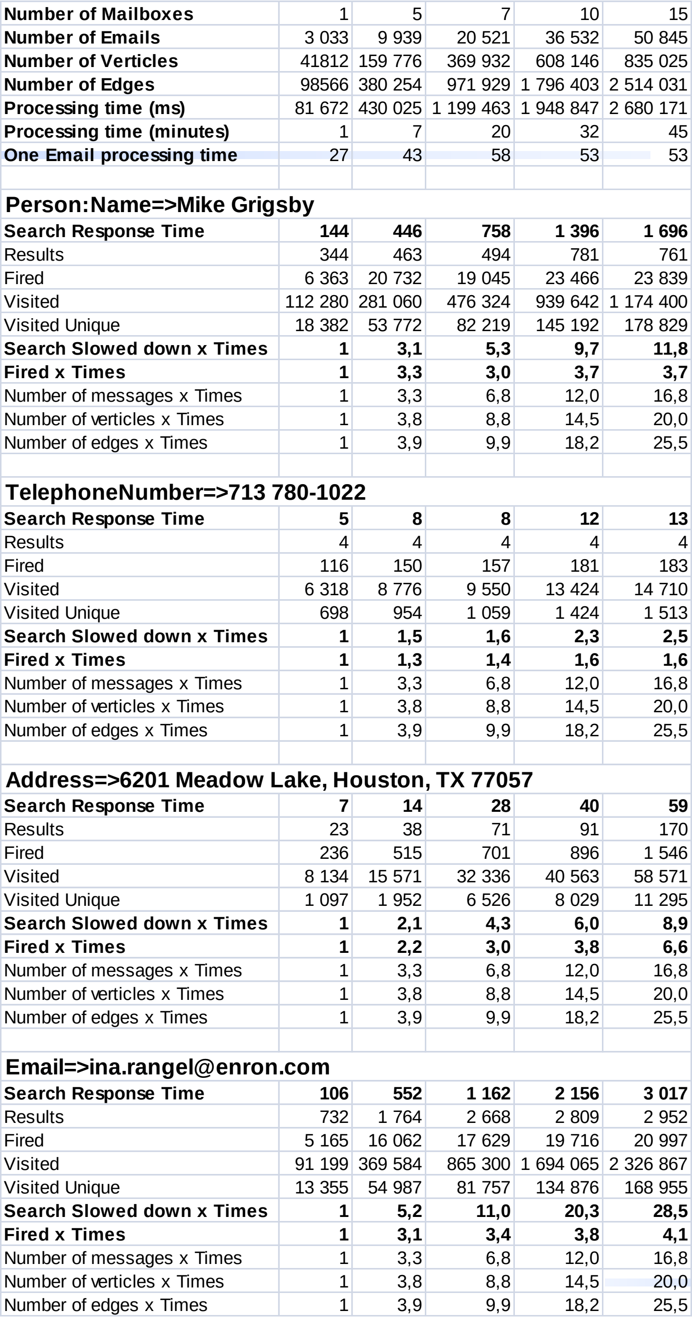
Algorithm and Evaluation



- All described in last year WI-IAT 2010 publication:
  - Laclavik et al. Use of Email Social Networks for Enterprise Benefit, IWCSN 2010
- Algorithm
  - breadth-first
  - Node fires only once
- Information Extraction Evaluation
  - Evaluation on set of 50 Spanish emails
  - Strict match 50-90%
  - Intersect match 80-94%
- Spread of Activation (relevance identification) Evaluation
  - 50 Spanish emails (phone/name):
  - Precision 60% (due to lower recall in IE)
  - Precision 85% (achievable with better IE)
  - self-healing (with new incoming emails)
  - 28 English emails: precision 77%

Performance evaluation

- Focus of this paper
- Experiments with 5 different sizes of dataset:
  - Number of visited nodes grows too high
  - Number of fired nodes grows acceptable
  - Search time ~ visited number of nodes
  - Scalability not possible with current implementation but achievable



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

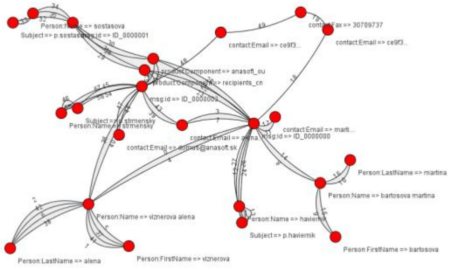
New Developments not included in the paper

- Faster algorithm
- Takes graph topology into account
- Breadth First
- Ends after it visit certain number of nodes (set to 10,000 experimentally)

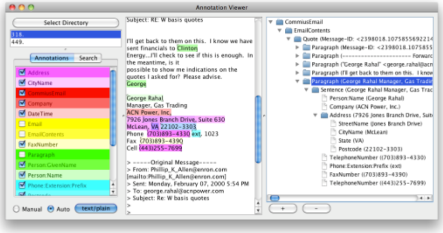
- Gives similar results as original algorithm
- Possibility for improvements:
  - It should take edge and vertex weight into account
  - Ignores multiple edges between nodes

```
private void computeRelatedBreadthFirst(Result start) {
    LinkedList<Result> rLL = new LinkedList<Result>();
    rLL.addLast(start);
    int count = visitNodeCount;
    rM.put(start, (double) count);
    vNodes++;
    while (!rLL.isEmpty() && count >= 0) {
        Result r = rLL.removeFirst();
        visited.add(r);
        int nCount =  g.g.getNeighborCount(r);
        double v = rM.get(r)/(double)nCount;
        if (v < threshold) //if value is to low we do not activate more
            continue;


        if (nCount<=count) {
            Collection<Result> rC = g.g.getNeighbors(r);
            for (Result result : rC) {
                if (!visited.contains(result)) {
                    rLL.addLast(result);
                }
                visited.add(result);
                double val = v;
                if (rM.containsKey(result))
                    val += rM.get(result);
                rM.put(result, val);
                vNodes++;
            }
            count -=nCount;
        }
    }
}
```

Conclusion


Person:Name => sostasova
Subject => p.sostasova
msg:id => ID_0000001
contact:Email => ce9f3...
contact:Fax => 30709737
contact:Email => ce9f3...
product:Component => anasoft_ou
product:Component => recipients_cn
msg:id => ID_0000002
Subject => p.strmensky
Person:Name => strmensky
contact:Email => alena...
msg:id => ID_0000000
contact:Email => marti...
contact:Email => domus@anasoft.sk
Person:LastName => martina
Person:Name => bartosova martina
Person:Name => viznerova alena
Person:Name => haviernik
Subject => p.haviernik
Person:FirstName => bartosova
Person:LastName => alena
Person:FirstName => viznerova

- Email Archives
  - Valuable source of knowledge
  - Hidden Social Networks owned by Enterprise or Individual
  - Information Extraction and Social Network Analysis can help
- Experiment
  - Pattern-based Information Extraction
  - Social Network Extractor
  - Spread of Activation
  - Scalable Relation identification with acceptable success rate



Subject: RE: W basis quotes

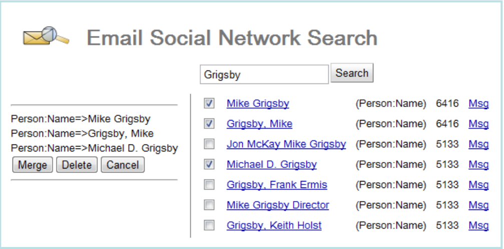
I'll get back to them on this. I know we have sent financials to Clinton Energy...I'll check to see if this is enough. In the meantime, is it possible to show me indications on the quotes I asked for? Please advise.
George

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
>
>

```
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 Tradi
        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)
```

- Applications
  - Recommendation and Search in Emails
  - Population of Databases (Cold start problem)
  - Possibility to extend social network graph with , processed document repositories and other business data
  - Business Intelligence and Knowledge Management



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