Semantic technology for capturing communication inside organization

Marko Grobelnik[1], Dunja Mladenić[1], Blaž Fortuna[1]
[1] Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
{Marko.Grobelnik, Dunja.Mladenic, Blaz.Fortuna}@ijs.si

Abstract. We address the problem of using semantic technology for capturing informal organizational structure by means of a light-weight ontology. The idea is to support knowledge management in analyzing communication between people in an organization. As an example we use social network of a mid size research institution obtained based on e-mail communication. We propose an approach consisting of five major steps that enable transformation of the data from a given e-mail transactions recordings to a light-weight ontology estimating structure of the organization. The experimental evaluation shows that on our e-mail transaction data the obtained organizational structure closely corresponds to the formal organizational structure. We conclude that the proposed approach is useful and applicable in real life situations where the goal is to model social structures based on communication records.

Keywords: knowledge management, organizational structure, knowledge discovery, semi-automatic ontology learning, e-mail transactions

1 Introduction

Knowledge management can be supported by capturing knowledge of people in an organization and also by capturing their communication records, such as, e-mail exchange. Semantic technologies have been successfully used in different domains, including medical care, legal support, government, homeland security, etc. where domain specific ontologies are developed and used for capturing knowledge and enabling reasoning in the specific domain. We propose an approach using semantic technology for capturing organizational behavior in the form of informal organizational structure obtained from social network of people in the organization. Output of the proposed approach is presented as a light-weight ontology.

In addition to relaying on semantic technology, the proposed approach incorporates usage of different knowledge discovery techniques for knowledge management [3] including data cleaning and transforming graph into a sparse matrix. We adapt the approach for semi-automatic ontology construction from textual data, where the candidate instances and classes for the ontology are lexical items each described by a

set of features. In the proposed approach we replace lexical items by nodes of the social network and describe each node by its context in the social network graph.

To evaluate the proposed approach we have performed experiments on real life dataset taken from a mid-size organization (700-800 people) and compared the obtained informal structure to the known, formal organizational structure. The dataset represents log files from organizational spam filter software giving us the set of e-mail transactions for the period of 19 months resulting in 2.7 millions of successful e-mail transactions used here for analysis and semi-automatic ontology learning.

The main contribution of this paper is adopting semantic technology to support knowledge management via analysis of social network [5] (in our case e-mail communication inside an organization). The approach consists from several steps that enable transformation of the people communication record to a light-weight ontology. Starting with log files from the institutional e-mail server we perform data cleaning and removing irrelevant e-mail transactions. The cleaned e-mail transactions are used to construct a graph where vertices are e-mail addresses that are connected if there is a transaction (an e-mail sent) between them. The e-mail graph is then transformed into a sparse matrix, where each e-mail address is represented as a sparse vector of features. Once having a set of sparse vectors, we apply an approach to semi-automated ontology construction as implemented in the OntoGen tool for semi-automatic, data driven ontology construction [2]. Semantic technologies and social network analysis both have some history; contribution of this paper is placed on adapting and combining existing technologies to support knowledge management, developing system and performing experiments on real-world data (e-mail graph as fairly popular application domain for discovering communities [20]).

The rest of this paper is organized as follows. Section 2 gives background of the proposed approach with related work. The approach is described in Section 3. Application of the proposed approach on real-world data of e-mail graph is described in Section 4. Experimental evaluation is given in Section 5. Some ideas for future work are described in Section 6.

2 Background and Related Work

The proposed approach combines several research results. It is based on the idea of using knowledge discovery techniques for knowledge management when pre-processing the data and using knowledge discovery techniques to support semantic technologies (Section 2.1). It further deals with analyzing communication between people in an organization, where the communication is presented with a social network (Section 2.2). The communication is modeled as a light-weight ontology using semantic technologies as integrated in a semi-automatic, data-driven approach (Section 2.3).

2.1 Knowledge Discovery Techniques for Knowledge Management

Knowledge discovery can be defined as a process which aims at the extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information from data in large databases [6]. Knowledge discovery techniques can be used to support semantic technologies along several dimension, as outlined in [3] combining methods from several research fields the most outstanding being Machine Learning and Data Mining [18, 6, 21,15], Statistics and statistical learning [16, 14], Information Retrieval [20], Natural Language Processing [17]. Knowledge discovery techniques are mainly concerned with discovering structure in data; this can serve as one of the key mechanisms for structuring knowledge. We commonly refer to such approaches as "ontology learning" which is usually performed in automatic or semi-automatic mode [8], where the goal is to extract the structure from data-sources into an ontological structure being further used in knowledge management process. Applications of semantic technologies are typically associated with more or less structured data such as text-documents and the corresponding metadata at some fix point in time. However, there are also applications in the areas where the data is in different form than text, such as, multimedia, signals or graphs/networks and, applications where the data and its semantic structure changes over time (eg., stream ontologies [7]). For most of such scenarios an extensive human involvement in building models from the data is not economical anymore. The advantage of automatic or semi-automatic methods offered by knowledge discovery technologies is even more evident than on textual data.

2.2 Social Network Analysis

Social network analysis [5] is gaining popularity with the growth of Web. Maybe one of the most prominent everyday usages of information coming from the network is ranking web search results which partly led to success of the Google search engine. Current research on social networks is centered on the problems such as, how to handle dynamics of networks; visualization of very large networks; creation of generative models to explain underlying laws of network generation; interchanging network data with other data modalities (such as text, images etc); and different applications on top of the network data, such as modeling the spread of an influence, modeling trust, improving search results, collaborative methods etc. Addressing dynamics of the networks is gaining popularity especially in connection to the Internet graph, modeling its evolution [23] and investigating different properties (power-law degree distribution [24], shrinkage diameter phenomena [25]). Dynamics of social network is commonly connected to study of community identification [26] and evolution [27, 28].

2.3 Social Network Analysis and Semantic Technologies

Semantic social networks is mentioned in the context of introducing a three-layered model which involves the network between people, the network between the

ontologies they use and a network between concepts occurring in these ontologies as proposed in [31]. However, what we are proposing here is starting with a social network and applying semantic technologies on it.

Semantic analytics was applied on social networks in [32], where a Semantic Web application is proposed that detects Conflict of Interest relationships among potential reviewers and authors of scientific papers. It is based on creating an ontology by merging a social network of co-authorship from a public bibliography database and a social network of friends from a social networking platform. As opposite to our work where we are identifying structure in a social network, the emphases in [32] is on merging the two social networks and finding if there is a conflict of interest relationship among people.

Our work goes into the category of problems where the goal is to find structure in large networks. More precisely, we aim at discovery of community structure within organizations based on e-mail log files. The closely related work [30] addresses the same problem using betweenness centrality to identify communities, while we propose to use semantic technologies to semi-automatically model the communities. Our motivation is in the direction of using the identified structure to gain some insides into the underlying processes which generated the network – this would allow us to detect and possibly explain phenomena within the network.

2.4 Semi-automatic Data-driven Ontology Construction

Semantic technologies as integrated in a semi-automatic, data-driven approach are based on combining different knowledge discovery techniques in the ontology learning framework. In this section we give a brief description of the approach originally proposed in [2] and implemented in OntoGen, a system for semi-automatic data-driven ontology construction that suggests concepts, relations and their names, automatically assigns instances to concepts and provides a good overview of the ontology to the user trough concept browsing and visualization. At the same time the user can fully adjust all the properties of the ontology by manually adding or deleting concepts, relations and reassigning instances. Most of the aid provided in the approach is based on some underlying data provided by the user at the beginning of the ontology construction. Data reflects domain of the ontology the user is building.

The central parts of the approach are the methods for discovering concepts from a collection of documents based on (1) Latent Semantic Indexing, as a well known technique for extraction of hidden semantic concepts or topics from text and (2) clustering, as a well established technique for partitioning the data based on some similarity measure. The approach incorporates two methods for suggesting concept naming, one providing descriptive keywords based on the concept cluster centroid and the other providing distinctive keywords based on Support Vector Machine model of the concept contrasting it to its neighboring concepts.

3 Approach Description

Traditional Semantic Web deals with ontologies constructed mainly from text documents. Especially ontology learning techniques deal almost exclusively with the problem of extracting and modeling the knowledge from text documents. The reason for this is that text is the most natural way of encoding information with the attached semantics. But text is not the only data modality which could be modeled using ontological structures. In this paper we propose an approach to building ontological models from social network data. In particular, we show an example of modeling internal communication of an organization (reflecting its organizational structure) from its e-mail server log files.

3.1 The main idea

Approaches to ontology learning from text could be summarized in the following main steps: (a) first, to extract candidates for future ontological instances and ontological classes, (b) next, describe them with appropriate features, and (c) finally with a set of heuristic rules or analytical approaches (e.g. similarity measures etc.) decide about the ontological elements and their connectivity. An important fact is that each element is described with a set of attributes which carry information of the element and which allow further processing. In the cases where we process text, this information usually tells us about the nature of the words or phrases, their context within the text etc.

For learning ontologies from social network we follow the same intuition as used for text. Since a social network is a general graph structure (vertices connected with directed or undirected edges), we redefine the above approach in the following way:

- instead of textual lexical elements (words and phrases) we use graph vertices (points in the graph);
- instead of features describing properties of lexical items we describe vertices with a context where they appear in the graph;
- instead of heuristic and analytical rules applicable on text we propose and develop new ones appropriate for dealing with graph data.

Apart from the above three replacements all other operations are shared between approaches for learning ontologies from text and from social networks.

3.2 Hypotheses and Contribution

There are two main hypotheses that we address. The first is that a communication graph (social network) data can be modeled by ontological structures (taxonomies). The second is that in a research institution e-mail data roughly correspond to organizational structure on an institution.

The main contributions of this work are the following. Based on knowledge discovery methods we propose a way for transforming social network data into the form suitable

for ontology learning (sparse vector representation). Using semantic technology we construct a light-weight ontology from organizational behavior as recorded in e-mail server log data. We evaluate the constructed model by comparing the informal organizational structure as evident form communication record to the formal organizational structure.

3.3 Main Steps

Our approach to ontological modeling of communication record of people consists of the following steps:

1. Obtain data recording communication activity. For instance, log files from the institutional e-mail server including information about e-mail transactions described by the following: time of the transaction, sender e-mail address and the list of receiver e-mail addresses.
2. Perform data cleaning including removing non-valid and non-relevant records, such as, removing non-valid e-mail transactions (no recipient, mailing lists, automatically generated e-mails, spam, etc.). This step can also include reducing the data to information relevant for further processing. For instance, removing time information from the e-mail log files to get the data which include e-mail addresses of sender and receiver.
3. From a set of e-mail transactions construct a graph where vertices are e-mail addresses and two vertices are connected if there exists an e-mail transaction between them. Similar as in [30], we only use sender and receiver information from the e-mail log files.
4. The e-mail graph is transformed into a sparse matrix, where *(i,j)* element of the matrix is non-zero if *i*-th and *j*-th vertices are connected directly or indirectly over the number of hops given as a parameter.
5. The sparse matrix representation of the graph is analyzed using semantic technology to produce an ontological structure roughly corresponding to the organizational structure of the institution where the e-mails came from.

Figure 1 shown the steps of our approach on an example, where we are applying the proposed approach for modeling informal organizational structure as recorded in e-mail log files.

The proposed approach consists of 5 steps but as the intermediate result is a matrix representing connections between the people, this can be also used by other approaches. The approach can be nicely complemented by applying social network analysis methods on the matrix obtained after the step 4. For instance, one could calculate centrality to identify the most central persons in an organization or, calculate network reach to get an idea how well connected are the people in an organization.

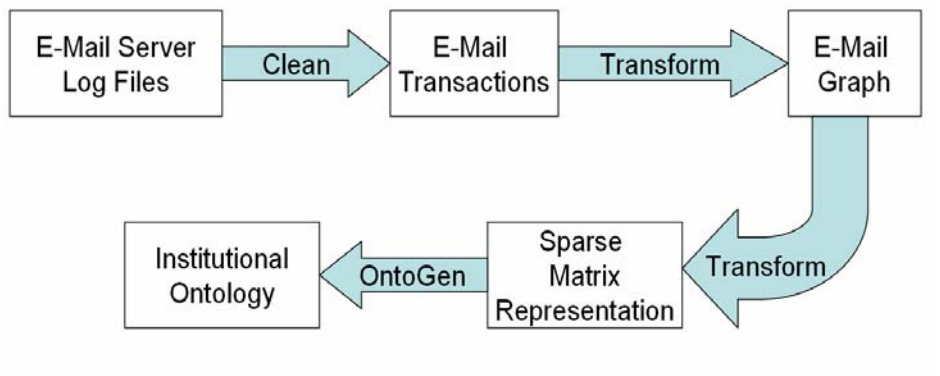


Fig. 1. Diagram showing transformation of the e-mail server log data into an ontological representation of institution.

4. Experiments

4.1 Data Description

The data used in our experiment is a collection of log files with e-mail transactions from amid size research institution. It was obtained from spam filter software. Each line of the log files denotes one event at the spam filter software. We were interested in the events on successful e-mail transactions which include information on time when the event happened, who sent the e-mail (sender), and who received it (a list of receivers). An example of a successful e-mail transaction is the following line:

```
2005 Mar 28 13:59:05 patsy amavis[33972]: (33972-01-3)
Passed CLEAN, [217.32.164.151] [193.113.30.29]
<john.nj.davies@bt.com> -> <marko.grobelnik@ijs.si>,
Message-ID:
<21DA6754A9238B48B92F39637EF307FD0D4781C8@i2km41-
ukdy.domain1.systemhost.net>, Hits: -1.668, 6389 ms
```

The above line tells us that there has been an e-mail sent from john.nj.davies@bt.com to marko.grobelnik@ijs.si at 2005 Mar 28 13:59:05. The advantage of using spam filter log files is additional information provided by spam filter software which tags each e-mail transaction as being “CLEAN” (the example above) or “SPAM”.

The log files include e-mails data from Sep 5th 2003 to Mar 28th 2005 which sums up to 12.8 Gb of data. After filtering out successful e-mail transactions we got 564Mb of data containing approx. 2.7 million of successful e-mail transitions. These 2.7 million of transactions were used for further processing. The whole dataset contains references to approx. 45,000 e-mail addresses. After the data cleaning phase, the

number is reduced to approx. 17,000 e-mail addresses out of which 770 e-mail addresses are internal inside the observed institution.

4.2 Data Cleaning

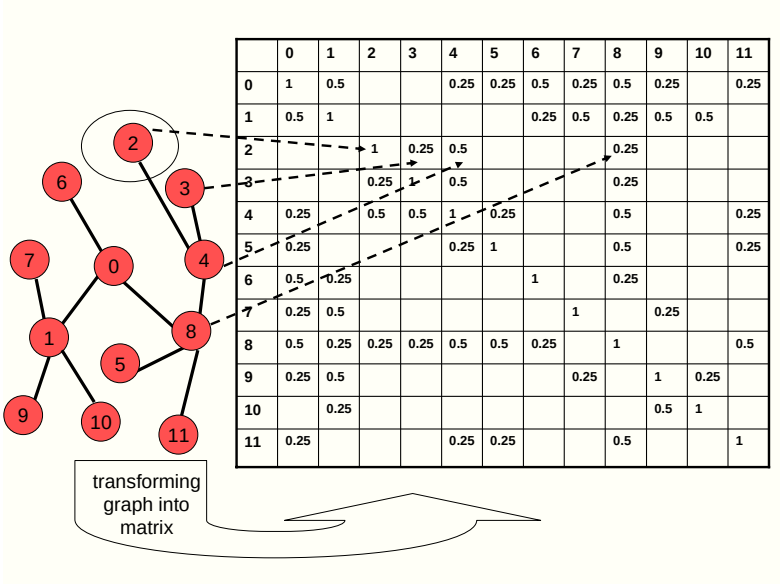
Spam filtering software log files records of all the successful e-mail transaction events – it doesn't control the existence of the e-mail addresses, successful delivery etc. It also includes all kinds of automatically generated email messages from mailing-lists, notifications etc. In general, for the analysis of institutional e-mails we are not interested in all the transactions which do not contribute to the overall goal of modeling the organizational structure in an institution. Therefore we perform three kind of data cleaning operations:

- Deleting all e-mail transactions which include e-mail addresses with escape and unusual characters. This operation deletes most of the automatically generated e-mail messages from e.g. notification or other similar services.
- Deleting all e-mail transactions where the pair <sender, receiver> appears less then a certain number of times (in our experiments we use the value 10). This filter deletes most the typos in e-mail addresses.
- Deleting all e-mail transactions with e-mail addresses which communicate with less then a certain number of different e-mail addresses (set to 10 in our experiments). This filter removes e-mail addresses which are rarely in use, contributing to removing potential noise and uninteresting transactions, as we are interested in frequent communications that characterize communities.

4.3 Data Transformation

In the process of processing the original log data towards ontological representation there are two significant data transformation we have applied (see Figure 1). Here we describe them in detail.

The first transformation transforms a collection of e-mail transactions into an e-mail graph. From the set of transactions we construct a graph where vertices are e-mail addresses and edges between vertices represent communication between the e-mail addresses. In other words, there is an edge between two vertices representing two e-mail addresses if there are e-mail transactions between them. Edges are additional labeled with the intensity of communication (number of transactions between e-mail addresses representing both vertices).



	0	1	2	3	4	5	6	7	8	9	10	11
0	1	0.5			0.25	0.25	0.5	0.25	0.5	0.25		0.25
1	0.5	1					0.25	0.5	0.25	0.5	0.5	
2			1	0.25	0.5				0.25			
3			0.25	1	0.5				0.25			
4	0.25		0.5	0.5	1	0.25			0.5			0.25
5	0.25				0.25	1			0.5			0.25
6	0.5	0.25					1		0.25			
7	0.25	0.5						1		0.25		
8	0.5	0.25	0.25	0.25	0.5	0.5	0.25		1			0.5
9	0.25	0.5						0.25		1	0.25	
10		0.25								0.5	1	
11	0.25				0.25	0.25			0.5			1

Fig. 2. Illustration of the graph transformation into a sparse matrix where the rows represent instances (vertices) and columns represent neighborhood with weights relative to the distance from the vertex in that row. Here we have set the maximal distance to d= 2. Notice that the diagonal elements have weight 1 (showing that the each vertex is in its own neighborhood). The dashed lines point out neighboring vertices and the corresponding weights for vertex labeled as 2. It has four non-zero elements in its sparse vector representation (1, 0.25, 0.5, 0.25) corresponding to four vertices (labeled in the graph as 2, 3, 4, 8).

The second transformation transforms an e-mail graph into a sparse matrix. Here we use a representational trick proposed in [4] which helps to compare and calculate similarity between vertices in the graph represented as rows of a sparse matrix. Figure 2 illustrates the graph transformation on an example graph assuming all the edges in the graph have weight 1. Matrix is formed following the formula proposed in [4] defining neighborhood of a vertex to contain all the vertices at the distance of up to *d* steps from the vertex. Consequently, non-zero components represent the neighbors at step *1, 2, 3, … d*. The values in the matrix are calculated using the formula $1/2^d$, where *d* is the distance in the number of steps from the vertex. This can be seen as a very rough but computationally efficient approximation of probabilities that a random walker on the graph, starting from the vertex [29], would reach a neighboring vertex normalized by the number of neighbors.

Similarity between vertices (i.e. rows in the sparse matrix) is calculated by using cosine similarity [19], a measure commonly used in information retrieval and text-mining when dealing with sparse vectors. Notice that different similarity measures are used on different tasks and for sparse vectors cosine similarity is preferred over Euclidean distance and distances used for string matching [22]. Having an e-mail graph represented in the form where we can efficiently compare vertices via similarity

of their vector representations, allows us to use most of the methods for unsupervised learning (clustering) which is the natural basis for ontology learning approaches.

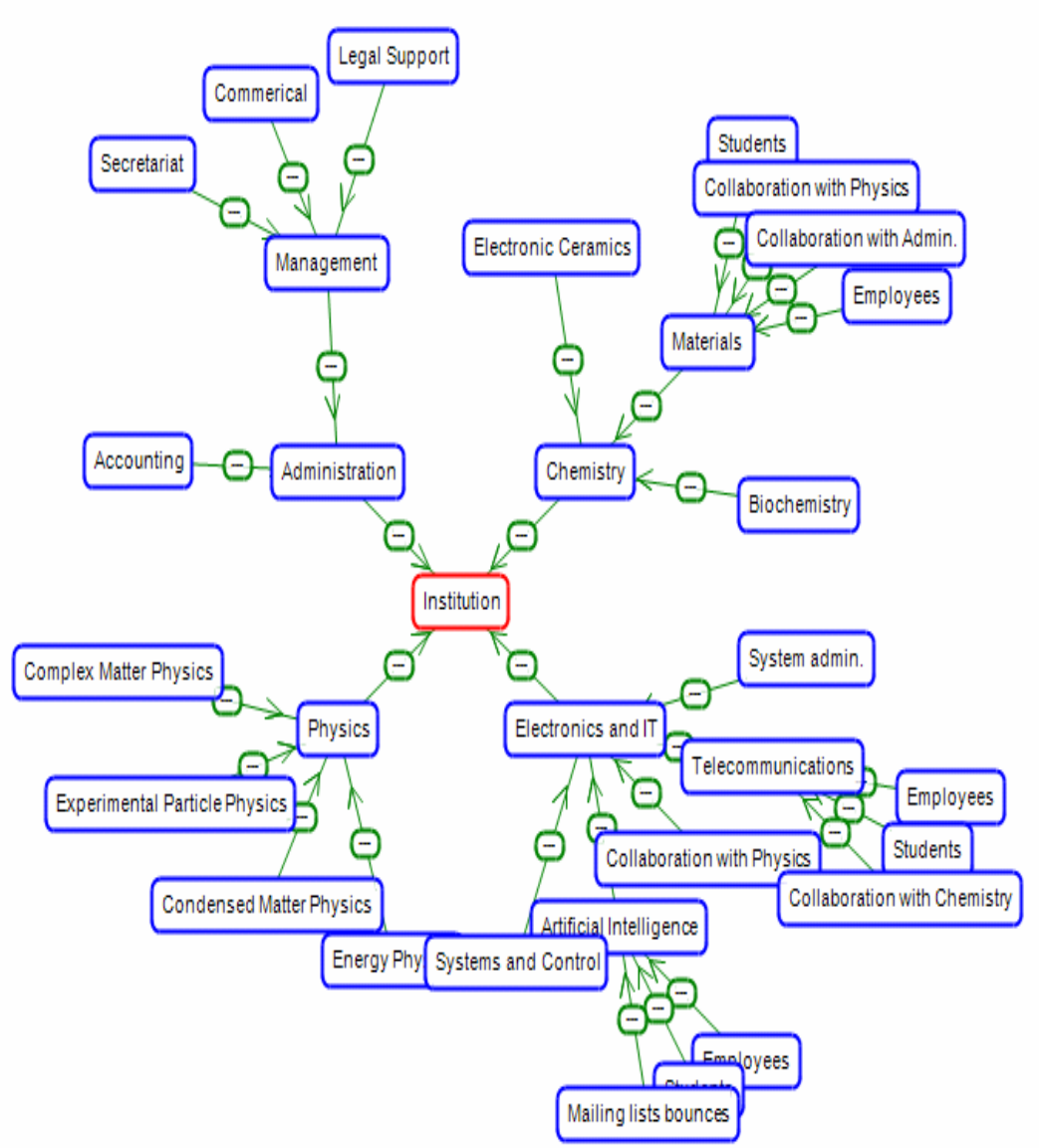


Fig. 3. Organizational structure modeled from the e-mail data in a 10 minute session with the ontology learning system OntoGen.

4.4 Ontology Modeling

For ontological modeling of the data we used the ontology learning system OntoGen [2], described in Section 2.4. The system can handle data represented as a set of feature vectors describing properties of ontological instances. Using several machine learning techniques OntoGen helps the user to construct an ontological structure directly from the data by suggesting sub-concepts (in our case sub-communities) and their naming, while letting the user to make final decisions. Input for ontology learning is a set of ontological instances (represented as sparse vectors) – in our case each instance corresponds to one e-mail address within the organization (one row from the matrix generated in step 4 in Section 3.3). On the output we want to get an ontology in the form of taxonomy modeling relation "subcommunity-of" which would correspond to the organizational structure of the institution where the e-mail data is coming from. The actual experiment consisted of 770 e-mail addresses from a mid size research institution. Each of these e-mail addresses was described with the subset of 17000 e-mail addresses being in the direct or indirect contact with the target e-mail address. Each e-mail address was represented as sparse vector from the e-mail graph.

The result of approx 20 minute session with OntoGen is shown in Figure 3. The background information that we have used there is associating e-mail addresses with organizational units. We can see that the whole organizational structure can on the top level be represented by four communities: electronics and information technologies, chemistry, physics and administration. As expected the administration consists of two sub-communities: management and accounting, where management can be further split into secretariat, legal and commercial support. However, at this point it is not clear if the administration is mainly communicating internally or fulfils its natural role of supporting other organizational units. From visualization in Section 4.5 we will see that indeed it fulfils its role of connecting the research groups. Inside each research group we can see sub-communities that are related to topics, such as, telecommunications, systems and control, artificial intelligence. On a finer level these topic oriented communities have their sub-communities that capture role of different people in the organization. For instance, artificial intelligence has a sub-community of PhD students, a sub-community of employees and a small sub-community getting mailing lists bounces. Closer inspection using OntoGen shows that this small sub-community consists of researchers also associated to some other institution (having different e-mail address there) sending e-mails to internal mailing lists that bounced (because they were sent from their e-mail address that belongs to the other institution).

An interesting hidden patter surfaces here showing a sub-community of physics is closely related to electronic and IT community (appears as its sub-community in Figure 3). Investigating status of that members several years later, we found that most of them became members of the same spin-off.

4.5 Visualization

One possible way of present complex data is to use techniques for visualization of high dimensional data. These typically include the major step of reducing high dimensionality of the observed data down to two or three dimensions while preserving the high level relationships between the data points. In the same way we can visualize the e-mail transactions data. For illustration we use here Document Atlas [1], a tool for visualization of document collections.

Figure 4 shows the anonymized dataset from the previous section in the geographical relief visualization. Each data point represents one e-mail address, e-mail addresses having similar communication are closer on the image and the density of points results in higher elevation of the terrain. From the image it can be seen the top level structure of the communication, corresponding to the top level concepts in the generated ontology. We can see three major "mountains" representing the major research topics of the instition functioning: computer science (electronics and IT), chemistry, physics. Management/administration is placed in the middle of the terrain serving as a connector between the three main research areas. It is interesting to notice that the institution management not only formally but also actually has its role in communication with all the other organizational units. On the other hand computer science has a rather pointed "mountain" showing intensive communication inside the computer science community.

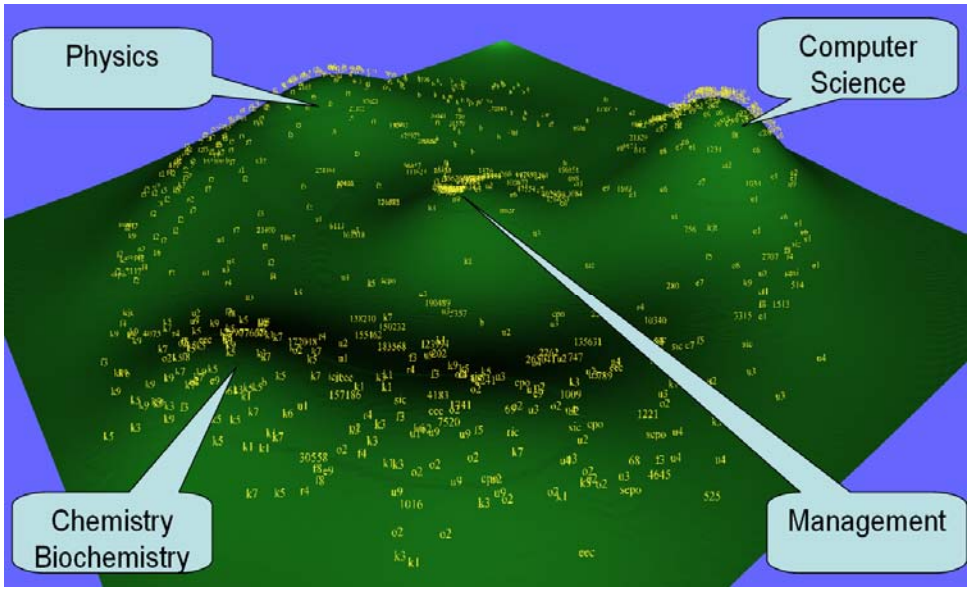


Fig. 4. Visualization of an e-mail communication record as geographical terrain. The high level structure of the data shows 4 major areas on the map corresponding to the four major groups, three research units (computer science, chemistry, physics) and management.

5 Evaluation

We perform evaluation of the approach described in the previous sections indirectly by showing the compactness of the clusters as produced by semi-automatic ontology construction tool OntoGen. The main hypothesis we want to prove is that communication intensity follows organizational structure of an organization – in other words, people inside the same organizational unit are communicating more intensively between each other then to the people outside their organizational unit. To show that, we perform a "golden standard" style of comparison, where we compare the ontological structure obtained from the communication record using the proposed approach to the formal organizational structure of the institution.

Table 1 shows the result of k-means clustering using 10 for the number of clusters (user parameter), as provided by OntoGen on the first ontology level. Columns correspond to the clusters and rows correspond to the organizational units of the institution. For each cluster we give the percent of e-mail addresses falling into the individual organizational unit – the sum of each column is 1. Analysis of the table shows that individual clusters obtained from the e-mail exchange graph actually contain e-mails belonging mainly to one of the formal groups. For instance, cluster `C-0` contains mainly e-mail addresses associated to organizational unit AI1 (artificial intelligence 1), `C-1` mainly form `SC (systems and control)`, `C-3` mainly from (Ce) `Ceramics`, etc..

Table 1. Part of the results for data grouped into 10 clusters (C-0, C-1, …C-9) showing distribution of the clustered e-mails over the formal groups inside the institution. The largest group in each automatically obtained cluster is marked with (!).

Group	C-0	C-1	C-2	C-3	C-4	C-5	C-6	C-7	C-8	C-9
AI1	**0.368!**				0.03					
Ch1			0.04	0.02			0.1	0.04	0.06	0.01
BC	0.21	0.05	0.01		0.03		**0.361!**	0.07	0.04	0.07
SA						**0.200!**			0.02	
SC		**0.381!**					0.01			
Ch2			0.14		0.06		0.01			
Ph					0.06		0.03	0.08		**0.261!**
Ch3	0.01	0.02	**0.235!**		0.03	0.1	0.01	0.01	0.06	0.02
Ce	0.01		0.01	**0.448!**		0.05	0.01			
MPh		0.03	0.03		**0.444!**	0.15	0.01	**0.347!**		0.17
AI2	0.15		0.01	0.02			0.01	0.01		
EPh			0.01		0.08		0.01	0.03	**0.471!**	0.01
…	…	…								

Looking more closely into cluster `C-0`, we can see that, 36.8% of all the e-mail addresses in this cluster are from `AI1(artificial intelligence), 21%`

of e-mail addresses are from `BC (Biochemistry)` and 15% are from `AI2 (artificial intelligence).` Knowing background of the institution organization, it is not surprising, as `AI1` and `AI2` used to be the same organizational unit shortly prior to our data collection and still have some common mailing lists and collaborations. The analysis suggests that community formed around this computer science community contains considerable proportion of members that belong to organizational unit covering biochemistry, which is a kind hidden pattern suggesting collaboration between them. After some discussions with the members of the identified community it turned out that indeed some researchers collaborate and work on bioinformatics.

Looking more closely into the identified communities, we can see that the most compact are clusters C-3, C-4 and C-8 where over 40% of all the communication is inside one formal organizational unit. On the other hand, C-2, C-5 and C-9 represent communities that connect different formal organizational units. As expected, if we increase the number of groups the compactness of clusters gets better.

6 Discussion and Future Work

The proposed approach can be seen as consisting of two main phases. The first phase on data pre-processing is an off-line task including data cleaning, graph construction and data transformation. The second phase is an interactive, exploratory analysis involving the user. It involves ontology modeling and data visualization. It is important to point out that the second phase enables not only semi-automatic construction of the community taxonomy but also exploring the data. This includes operations such as, zoom into details of visualization, inspect details of the communities all the way to reading the content (e-mail address in our case), decide to exclude some of the instances from communities, split community further down into smaller sub-communities, etc. In the paper we have presented only one of the possible images obtained by each of the two used system. Notice that both systems are publicly available and enable interactive usage on different data [1, 2].

In the future work we plan to address dynamic component of the e-mail data. This would give us insight into dynamic nature of a life of an organization, it would allow us to better understand social phenomena (such as getting new project, employing new person, estimating social value of a person etc.) as well as trying to predict future phenomena based on the past experience.

We would also like to include additional knowledge and contextual information about the organization, types of people etc. This would give us possibility to model relations between the people and detect types of people. For instance, it would be possible to make a model of a secretary and associate other people in the organization with this model to see how the communication patterns follow certain profile.

Acknowledgments. This work was supported by the Slovenian Research Agency and the IST Programme of the European Community under SEKT Semantically Enabled Knowledge Technologies (IST-1-506826-IP), NeOn Lifecycle Support for Networked Ontologies (IST-4-027595-IP) and PASCAL Network of Excellence (IST-2002-506778).

References

1. Fortuna, B., Mladenic, D., Grobelnik, M. Visualization of text document corpus. *Informatica journal* (Ljubljana), 2005, vol. 29, no. 4, pp. 497-502. (tool available at http://docatlas.ijs.si)
2. Fortuna, B., Grobelnik, M., Mladenic, D. Semi-automatic construction of topic ontology. In *Semantics, Web and Mining* (Berendt et al eds.), EWMF 2005 and KDO 2005: Revised Selected Papers, Lecture Notes in Artificial Intelligence, Springer 2006, (tool available http://ontogen.ijs.si).
3. Grobelnik, M., Mladenic, D. Automated knowledge discovery in advanced knowledge management. *Journal of knowledge management*, 2005, 9:132-149.
4. Mladenic, D., Grobelnik, M. Visualizing very large graphs using clustering neighborhoods. In: Morik et al (eds.), Local pattern detection : international seminar : Dagstuhl Castle, Germany, April 12-16, 2004 : revised selected papers, Lecture notes in computer science, Lecture notes in artificial intelligence, 3539, State-of-the-art survey. Berlin; Heidelberg; New York: Springer, cop. 2005, 89-97.
5. Wasserman, S., Faust, K. Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences), Cambridge University Press, New York, 1994.
6. Fayyad, U., Piatetski-Shapiro, G., Smith, P., and Uthurusamy R. (eds.). Advances in Knowledge Discovery and Data Mining. MIT Press, Cambridge, MA, 1996.
7. Grobelnik, M., Brank, J., Mladenic, D., Novak, B., Fortuna, B. Using DMoz for constructing ontology from data stream. Proceedings of the 28th International Conference on Information Technology Interfaces ITI-2006, June 19-22, 2006, Cavtat/Dubrovnik, Croatia, (IEEE Catalog, No. 06EX1244). Zagreb: University of Zagreb, SRCE University Computing Centre, cop. 2006, 439-444, 2006.
8. Grobelnik, M., Mladenic, D. Knowledge discovery for ontology construction. In: Davies, Studer, Warren (eds.), *Semantic web technologies : trends and research in ontology-based systems*. Chichester: John Wiley & Sons, cop. 2006, 9-27.
9. Brank, J., Grobelnik, M., Mladenić, D. *A Survey of Ontology Evaluation Techniques*. In Proceedings of SiKDD-2005 Conference, Ljubljana, Slovenia, 2005.
10. Ehrig, M., Haase, P., Stojanovic, N. *Similarity for ontologies - a comprehensive framework*. In Workshop Enterprise Modeling and Ontology: Ingredients for Interoperability, 5th International Conference on Practical Aspects of Knowledge Management, Vienna, Austria, 2004.
11. Fortuna, B., Grobelnik, M., Mladenic, D. Background knowledge for ontology construction. In Proceedings of ECAI-2006 Workshop on Contexts and ontologies: theory, practice and applications, C&O-2006, Trentino, Riva del Garda, Italy, 28 August - 1 September 2006, Trento: ITC-IRST, 2006, 74-80.
12. Salton, G. *Developments in Automatic Text Retrieval*. Science, Vol 253, 974-979, 1991.
13. Grobelnik, M., Mladenic, D., Simple classification into large topic ontology of web documents. CIT. Journal of Comput. Inf. Technol., 2005, 13:279-285.
14. Duda, R. O., Hart, P. E. and Stork, D. G. (2000). Pattern Classification 2nd edition, Wiley-Interscience.

15. Hand, D.J., Mannila, H., Smyth, P. Principles of Data Mining (Adaptive Computation and Machine Learning), MIT Press, 2001.
16. Hastie, T., Tibshirani, R. and Friedman, J. H. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Series in Statistics, Springer Verlag, 2001.
17. Manning, C.D., Schutze, H. Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge, MA, 2001.
18. Mitchell, T.M. Machine Learning. The McGraw-Hill Companies, Inc., 1997
19. Rijsbergen, C. J., van.Information Retrieval, Butterworths, 1979.
20. Tyler, J. R.,Wilkinson, D. M., and Huberman, B. A. Email as spectroscopy: automated discovery of community structure within organizations. In International Conference on Communities and Technologies, 81-96, Deventer, The Netherlands, Kluwer, 2003.
21. Witten, I.H., Frank, E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 1999.
22. Cohen W.W., Ravikumar P. and Fienberg S.E. A comparison of string distance metrics for name-matching tasks. In Proceedings of IJCAI03 Workshop on Information Integration on the Web (IIWeb03), Acapulco, Mexico, 2003.
23. Albert, R., Barabasi, A. L.. Emergence of scaling in random networks. Science, 1999.
24. Faloutsos, M. , Faloutsos, P. , Faloutsos, C. On power-law relationships of the internet topology. In SIGCOMM, 1999.
25. Leskovec, J., Kleinberg, J., Faloutsos, C. Graphs over time: Densification laws, shrinking diameters and possible explanations. In Proceedings of 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2005.
26. Kumar, R,. Raghavan, P., Rajagopalan, S., Tomkins A. Trawling the Web for emerging cyber-communities. In Proceedings of the Eighth World Wide Web Conference, 1999.
27. Backstrom, L., Huttenlocher, D. P., Kleinberg, J. M., Lan, X. Group formation in large social networks: membership, growth, and evolution. In Proceedings of 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006.
28. Aggarwal, C., Yu, P. Online analysis of community evolution in data streams. In Proceedings of ACM SIAM on Data Mining 2005.
29. Olston, C., Chi, H. E. ScentTrails: Integrating browsing and searching on the Web. In ACM Transactions on Computer-Human Interaction (TOCHI), Volume 10, Issue 3, Pages: 177–197, ACM Press, New York, NY, USA. 2003.
30. Tyler, J. R., Wilkinson, D. M., Huberman, B. A. Email as Spectroscopy: Automated Discovery of Community Structure within Organizations. In Communities and Technologies, Huysman, M., Wenger, E., Wulf, V. (eds.). Springer 2003.
31 Jung, J. J., Euzenat, J., Towards Semantic Social Networks, In The Semantic Web: Research and Applications, Proceedings of the 4th European Semantic Web Conference, ESWC 2007, Franconi, E., Kifer, M., May, W., (eds.). pp. 287 - 280, Lecture Notes in Computer Science, Springer 2007.
32. Aleman-Meza, B., Nagarajan, M., Ramakrishnan, C., Ding, L., Kolari, P., Sheth, A. P., Arpinar, I. B. Joshi, A., Finin, T., Semantic analytics on social networks: experiences in addressing the problem of conflict of interest detection, In Proceedings of the 15th international conference on World Wide Web, pp. 407-416, ACM, 2006.