Adding Taxonomies Obtained by Content Clustering to Semantic Social Network Analysis

Hauke Fuehres[1], Kai Fischbach[2], Peter A. Gloor[1], Jonas Krauss[1], and Stefan Nann[1]

[1] Center for Collective Intelligence MIT, Cambridge, MA, USA
[2] Department of Information Systems and Information Management, University of Cologne

Abstract. This paper introduces a novel method to analyze the content of communication in social networks. Content clustering methods are used to extract a taxonomy of concepts from each analyzed communication archive. Those taxonomies are hierarchical categorizations of the concepts discussed in the analyzed communication archives. Concepts are based on terms extracted from the communication's content. The resulting taxonomy provides insights into the communication not possible through conventional social network analysis.

1 Introduction

People increasingly publish information on their social networks on the Internet and on social network sites in special [1]. Various sources can be used to obtain the data on relations between different actors. Email archives as well as publicly available online-forums may serve, among others, as the sources of data to be analyzed [2]. Those interaction networks can be studied through measures of *social network analysis (SNA)*. Analyzing social networks with these measures reveals the structure of the networks. The information needed to model the network is often explicitly given or can easily be obtained.

The aim of this work is to combine the analysis of (electronically available) communication structures by means of social network analysis with the analysis of communication content by means of *information retrieval (IR)* and to introduce a software tool performing these tasks. This tool has been implemented as a module of the Condor toolkit. The content of communication sent in social networks is analyzed using information retrieval. The topics discussed in those communication messages are extracted and visualized. The automated construction of a taxonomy from the extracted topics helps to understand the relationships between them.

In contrast to the data representing the network structure, the content of the communication is usually unstructured. Interactions in the surveyed networks are often unstructured documents sent from one actor to one or more other actors, sometimes enriched with additional attributes like a timestamp. IR methods help to analyze the unstructured message content. Essential to those methods, like flat- or hierarchical clustering, is the introduction of a similarity measure on the words used in the communication content.

The idea behind the proposed approach is twofold: On one hand, the formal analysis of the communication structure by SNA methods is enriched with information on the communication content. Supplementary to the key insights gained by SNA methods, statements on the topics discussed by certain actors can be made. On the other hand, satisfying information needs using information retrieval can be supported by the formal knowledge on the network structure. SNA offers methods to assess the centrality of each actor in a network. Those key figures of actors are used to weight the information retrieved from messages they sent. Extraction of topics from the messages should not only be based on the topics' importance in the "flat text-file" but also show the context they are used in. A person's information acquisition is influenced by their social network [3]. Weighting in key data on those networks into the information retrieval process reflects the importance of those people's role in the information acquisition and diffusion process.

2 Related Work

Although the role of social network structures in document corpora is well known and utilized for information retrieval tasks, surprisingly little work deals with the extraction of term relations from the content of social networks. Usually the social network structure of documents is used to improve the document weights. Those weights modify the sorting order of retrieved documents in a search engine. A typical example of such a social network structure to be utilized to improve document ranking is the hypertext structure of web documents [4].

When focussing on the extraction of terms and leaving out the SNA component different information retrieval approaches can be identified. Supervised learning methods like the support vector machine are a prominent example [5]. This classifying technique was adapted to diverse purposes and requirements. One enhancement to the SVM relevant to the scope of this paper is the hierarchical support vector machine proposed in [6]. Instead of constructing a "flat classification", this technique allows to create hierarchical taxonomies.

Another important way of extracting topics from text corpora can be achieved by utilizing *latent semantic indexing (LSI)* [7]. It is an enhancement to the vector space model described in [8]. LSI is based on a singular value decomposition (SVD) of the term-document matrix. After calculating the SVD, the term-document matrix can be approximated with a lower-rank matrix. With this step a dimension reduction of the term-document vector space can be achieved. Instead of using the term-document vector space, IR methods can work on the reduced concept space defined by the SVD. The clustering algorithms applied to the problem of finding taxonomies of terms can work on the reduced concept space instead of using the original vector space. But more important to the goals of this paper is the ability to reveal hidden structures in the original vector space. Those hidden structures to be uncovered are polysemes and synonyms.

Having access to this information on terms may help to improve the quality of taxonomies obtained by clustering algorithms. Preprocessing with a stemmer can also be avoided since LSI will identify words with similar meanings as related concepts.

3 Semantic Social Network Analysis

SNA provides methods to analyze the interactions and relationships between actors in a network. The field of SNA emerged in behavioral science where interactions of people were analyzed. At the beginning there was the insight that interactions between individuals have influence on the individuals themselves [9, ch. 2]. Analyzing different relations between actors has been applied to different fields of study [10]. The methods have been successfully applied to Organisational Behavior [11] and the analysis of the spread of diseases [12].

Semantic social network analysis factors in content analysis of the relational data into the analysis of social networks. Therefore, it can be applied to social network data where interaction takes place by exchanging textual information. These can be found in email archives and networks built of websites and their linkage among each other. Another prominent example of textual interaction shared in a social network are the messages exchanged in online forums.

Semantic social network analysis as introduced in [13] allows to analyze the content of textual interaction in social networks together with the network structure. Techniques from information retrieval are used to extract important terms from the interaction's content.

Analyzing networks can be conducted in a static or a dynamic way. Traditional SNA focuses on a static view on the available data. The key measures used in SNA reflect a social network in a static way. However, networks analyzed with the means of SNA can be of dynamic nature. Networks might evolve over time. Identifying and understanding patterns in an evolving network can help to understand the nature of the whole network [14]. One possibility to gain insight into the dynamic structure of a network is to divide the data on the social network into several timeframes. The next step is to calculate SNA key measures for each of those timeframes and compare them over time [15].

4 Our Approach: Clustering

The goal of this work is to extract a *taxonomy* of terms and concepts from the interaction's content. This taxonomy should give an overview of the discussed topics in the content of the interaction in the analyzed social network. The elements to be categorized are the terms extracted from the interaction's content. A taxonomy created on top of those terms should help to understand the most important topics discussed in the interactions between actors of the analyzed social networks. Different ways of obtaining a (hierarchical) classification of a set of discriminable objects are known. In this section clustering methods are introduced. Generally, these methods are used to assign discriminable objects into groups. Clustering methods are applied in different fields of research. Those methods are used to group unlabeled data. Although the underlying idea of clustering methods is identical in all those fields, various terminologies and assumptions emerged [16]. In information retrieval clustering is often used to locate information. Using clustering to find information is used for a long time in libraries where books are classified by the topic they discuss [17].

Besides calculating the similarity of terms by utilizing their distribution among the analyzed documents, the similarity of terms can also be obtained from external sources. The idea behind this approach is to measure the similarity of terms in the analyzed documents by obtaining the semantic similarity of those terms from an existing taxonomy. Those taxonomies, or generally speaking lexical networks, can be obtained from different kinds of sources. The way on how to calculate the semantic similarity between two words might differ depending on the source of the lexical network used [18]. However, the similarity measures can be categorized into two different approaches [19]. The first category combines edge counting based methods whereas the second category roots in information theory based methods. Both approaches are described later on in this section.

The base of the similarity measures described in this section are lexical networks or lexical taxonomies. A lexical taxonomy is a tree-like structure with its nodes representing concepts. One source of background knowledge for measuring semantic similarity of words is the Wikipedia online encyclopedia. Wikipedia is an encyclopedia built on user generated content. It has a general scope with more than 3,222,261 articles in the English Wikipedia.[3] In Wikipedia authors are encouraged to add existing or new pages to categories and create new categories when necessary. The categories are arranged in a category network with a tree-like structure. This network of categories can be used to derive a semantic taxonomy. Several ways of extracting semantic taxonomies from Wikipedia are known. In [20] Wikipedia categories are used to identify topics of documents by relating the documents' content to category titles and to the titles of articles in categories. In [21] the structure of the Wikipedia's category network as well as the titles of the categories are used to extract semantic relations between different concepts.

4.1 Edge Counting Based Similarity

The first family of semantic similarity measures are edge counting based measures. Those similarity measures use the number of edges between two concepts in the graph representing the semantic network to calculate the similarity of those concepts. In [19] the basis of the edge counting measures is seen in [22]. A simple approach to calculate the similarity of two concepts is to use the path length of the shortest path from one concept to the other as the measure of similarity [23].

[3] Number from http://en.wikipedia.org/wiki/Special:Statistics in March 2010.

4.2 Information Based Similarity

Instead of using the edge counting similarity measures described in the previous section, an information based approach is introduced in [23] to calculate the distance between different semantic concepts. How much information two concepts share in common is the intuition of this similarity measure. The idea behind this approach is based on the information theoretical notion of information content of a concept. For each concept c in the taxonomy, $p(c)$ is the probability of encountering an instance of concept c. The higher a concept c_i is placed in the taxonomic tree the higher is its probability $p(c_i)$. If the taxonomy has one single root node its probability is 1. The *information content* of a concept c is $ic(c) = -\log p(c)$ and serves as the foundation for the calculation of the information based similarity. This information content decreases with the increasing of its probability. That means the more general a concept is the lower is its information content. The concept embodied by a single root node in a semantic taxonomy therefore has an information content of 0. In order to gain specific values for the probabilities of the concepts the frequencies of words in natural language corpuses can be used. The similarity of concepts based on the information content is defined by:

$$sim_{RES}(c_i, c_j) = \max_{c \in S(c_i, c_j)} -\log p(c) \qquad (1)$$

with $S(c_i, c_j)$ being the set of concepts subsuming both c_i and c_j[23]. With this definition the similarity of two concepts c_i and c_j in a semantic taxonomy is measured by the information content of the lowest common subsumer (LCS) of c_i and c_j. The lowest common subsumer of c_i and c_j is the lowest node in the semantic taxonomy that subsumes concepts c_i and c_j and thus is a hypernym of both concepts.

A notable generalization of information based similarity measures was introduced in [24]. The aim was a universal and theoretically justified model of similarity. Whereas other measures are bound to a particular domain or application, Lin's measure is only based on information theory. This omits assumptions based on the underlying domain. The definition of Lin's similarity is rooted in assumptions on the concept of similarity not in any specific formula. Different similarity measures for specific domains can be derived from those assumption. The derived semantic similarity is similar to Resnik's similarity measure:

$$sim_{LIN}(c_i, c_j) = \frac{2 * \log\ p(LCS(c_i, c_j))}{\log p(c_i) + \log p(c_j)} \qquad (2)$$

with $LCS(c_i, c_j)$ defining the lowest common subsumer of c_i and c_j.

4.3 Boosting Similarity of Terms with the Betweenness Centrality of the Actors

Although the corpuses analyzed with the introduced implementation can be of different nature they all have a social network structure in common. Social network analysis provides information on the structure of a network and on the position of actors in such a network. One powerful tool to assess the importance of an actor in a social network is the betweenness centrality of each actor. This measure reveals an actor's degree of centrality in a social network. Centrality thus can be interpreted as the importance of an actor. A message send by an actor can be linked to the importance of the sending actor. Thus a message can be weighted with the centrality of its sending actor.

Messages of less important actors can now be identified. In this way it possible to take only messages of important actors into account. The basis of analyzing terms and their similarity in this work is built on the vector space model. In a term-document vector space the distinction between different levels of importance of documents can be used for pruning the dimensionality of the vector space. Documents with a low importance can be ignored in the following analysis.

Besides reducing the dimensionality of the term-document vector space the importance weights of the documents can be used to calculate an importance weight for each term. Such an importance weight for a term is based on the importance weights of the documents and therefore is based on the betweenness centrality of the actors in the social network.

The hierarchical methods can work either by pooling all objects into one cluster and splitting up this cluster recursively or by starting with single objects and merging them into clusters. *Top-down clustering*, although less frequently used, has some advantages over merging algorithms. It is possible to stop the calculation when the clusters are fine-grained enough. Also, the global distribution of objects to cluster is taken into account [25, p. 396]. When using bottom-up or *hierarchical agglomerative clustering (HAC)* algorithms all decisions in the clustering process are made on a local basis without taking the global distribution into account. On the other hand, the top-down algorithms are more complex since flat clustering techniques are applied for each cluster to be split. In each step of an HAC algorithm the most similar clusters are merged. This procedure iterates till only one cluster is remaining that holds all terms. Alternatively the algorithm might be designed to stop when a certain number of top-level clusters remain. The more common hierarchical agglomerative clustering algorithms are discussed in the following sections and are the foundation of the described implementation.

Complete-Link Clustering The clustering algorithms depend on similarity measures on the terms to cluster. The introduced similarity measures are defined as functions $sim : \mathbb{T} \times \mathbb{T} \rightarrow [0, 1]$ with $\mathbb{T}$ being the space of terms to be compared. Since these functions are defined on the binary Cartesian product of the term vector space, new similarity measures are needed for comparing similarity of clusters of terms. Such a similarity measure of clusters needs to compare more than two terms with each other. These similarity measures on clusters yield the clusters to merge in each step by determining the most similar clusters.

A simple way of calculating the similarity of two clusters is *single-link clustering*. In each step, this algorithm merges the clusters with the nearest neighboring members. The single-link clustering is a local criterion since only one singleton member of each cluster is relevant to the calculation of the similarity. The similarity of the most similar members is the similarity of both clusters. The single-link clustering was introduced in [26]. Another way of calculating the similarity of clusters is the *complete-link clustering* algorithm. Instead of only paying attention to the most similar members of two clusters to calculate the cluster similarity, the diameter of the merged cluster is the crucial measure. The similarity of two clusters is the diameter of the merged cluster. This similarity can be calculated by assigning the similarity of the two most dissimilar singletons. In contrast to the single-link clustering the complete-link clustering is not local since the diameter of the whole cluster is taken into account. Therefore, the resulting clusters are more compact; clusters with small diameters are preferred by this method.

Group-Average Agglomerative Clustering The complete-link clustering introduced in the previous section chooses only one representing member of each cluster to calculate the similarity with the other clusters. Even though the diameter of each cluster is taken into account by the complete-link clustering, obstacles like the sensitivity against outliers persist. The *group-average agglomerative clustering (GAAC)* method uses the similarity of each member of the clusters to calculate the similarity of clusters. The aim of the GAAC algorithm is to build compact clusters. The average of the pairwise similarities of all members of both clusters is calculated [25]. It is important to mention that the similarity of members already in the same cluster is also taken into account.

By using the GAAC algorithm, the behavior of single-link clustering algorithms to create chains of clusters is avoided as well as the strong sensitivity towards outliers of complete-link clustering.

Since each member of each cluster is factored in in each step of calculating similarities and merging the most similar clusters, the time complexity can not be reduced with priority queues, as it is possible for single-link and complete-link algorithms. Thus the time complexity of the GAAC algorithm is in $\mathcal{O}(N^3)$. [27] shows how the complexity can be reduced to $\mathcal{O}(N^2)$, although with several constraints. This simplification only holds when the objects to cluster are represented by vectors in $\mathbb{R}^N$ and the applied similarity measure is the dot product. The key to the simpler calculation of the similarities of cluster C_i and cluster C_j is the definition of cluster similarity by: [25, p. 389]

$$sim_{GAAC}(C_i, C_j) = Norm(C_i, C_j) \left[(\sum_{t_m \in C_i \cup C_j} t_m)^2 - (|C_i| + |C_j|) \right] \quad (3)$$

with:

$$Norm(C_i, C_j) = \frac{1}{(|C_i| + |C_j|)(|C_i| + |C_j| - 1)} \quad (4)$$

5 Implementation and Evaluation

In this implementation the semantic social network analysis [13] module of Condor, the successor of TecFlow [28], is extended. Condor analyzes and visualizes communication and interaction networks. The data representing the network structure needs to be available in electronic format. Condor natively supports different data sources. Email archives can be imported from Eudora, Microsoft Outlook, or directly from an IMAP server. Weblinks and blogs can be accessed via Google's blogsearch and Microsoft's live search API. Data gathered with Social Badges can also be loaded into Condor [29]. Not natively supported data sources can be loaded into Condor via flat files or by parsing the data directly into a MySQL database. In extension to a static analysis of interaction networks, Condor and its predecessors can be used to study dynamic networks and their evolution over time. Monitoring social networks over time helps understanding the evolution of relationships in the networks [15]. Condor visualizes the social networks with a spring-embedder model developed by Fruchterman and Reingold [30]. This algorithm enhances Eades method [31] to places the nodes and the edges of a network on a two-dimensional plane. The screenshot in figure 3 shows the resulting structure of the force-directed algorithm.

An important feature of Condor is its ability to process and visualize temporal information on social networks. Especially when analyzing the content of the communication the temporal distribution of terms used by actors in the network is of interest. An increased use of terms pooled in certain clusters at one moment in time could point to the topics discussed in the social network during that time. To support users in assessing the prominence of single nodes in the taxonomies and the concepts these nodes represent this implementation allows to view the temporal distribution of each node in a chart. Figure 1 shows an example of such a distribution chart.

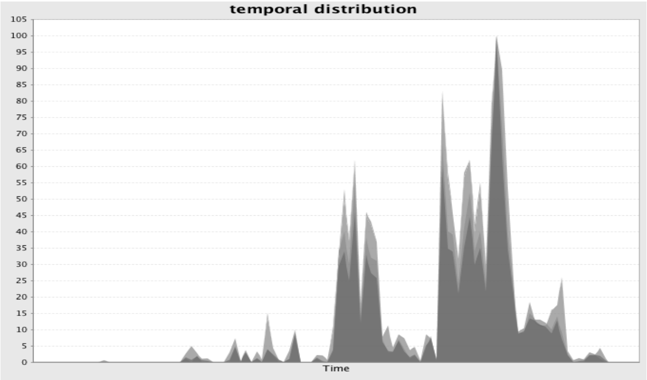
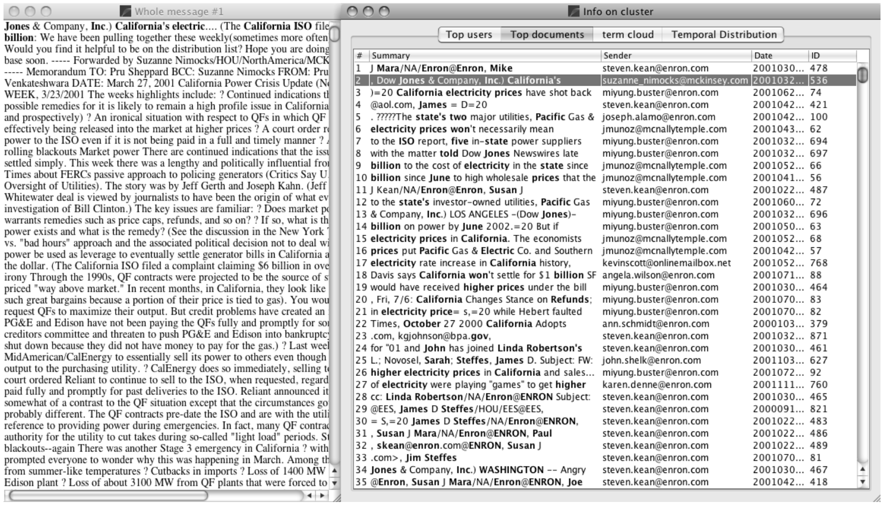


Fig. 1. Screenshot of the temporal distribtion chart.

In order to contrast the cluster scrutinized by the user to all the other clusters in the taxonomy the distribution of all terms is shown in each chart too. Values used in the temporal distribution chart are weighted with the terms' importance scores as described in section 4.3. By using those importance scores instead of the bare numbers of term occurrences, the betweenness centrality of the actors using the terms is factored in. Utilizing the betweenness based importance of each term punishes those terms and concepts used by less important actors.

An additional view on the data in each cluster is the list of the most important documents shown in figure 2. For each document a summary as well as the sender's name and the submission date is given. The documents are ordered by their importance for the terms in the selected cluster. On the left side of figure 2 an additional window with the content of the selected document can be seen.



Jones & Company, **Inc**.) **California's electric**.... (The **California ISO** file
billion: We have been pulling together these weekly(sometimes more often
Would you find it helpful to be on the distribution list? Hope you are doing
base soon. ----- Forwarded by Suzanne Nimocks/HOU/NorthAmerica/MCK
----- Memorandum TO: Pru Sheppard BCC: Suzanne Nimocks FROM: Pru
Venkateshwara DATE: March 27, 2001 California Power Crisis Update (N
WEEK, 3/23/2001 The weeks highlights include: ? Continued indications th
possible remedies for it is likely to remain a high profile issue in California
and prospectively) ? An ironical situation with respect to QFs in which QF
effectively being released into the market at higher prices ? A court order r
power to the ISO even if it is not being paid in a full and timely manner ?
rolling blackouts Market power There are continued indications that the issu
settled simply. This week there was a lengthy and politically influential fro
Times about FERCs passive approach to policing generators (Critics Say U
Oversight of Utilities). The story was by Jeff Gerth and Joseph Kahn. (Jeff
Whitewater deal is viewed by journalists to have been the origin of what ev
investigation of Bill Clinton.) The key issues are familiar: ? Does market p
warrants remedies such as price caps, refunds, and so on? ? If so, what is th
power exists and what is the remedy? (See the discussion in the New York
vs. "bad hours" approach and the associated political decision not to deal wi
power be used as leverage to eventually settle generator bills in California a
the dollar. (The California ISO filed a complaint claiming $6 billion in ove
irony Through the 1990s, QF contracts were projected to be the source of s
priced "way above market." In recent months, in California, they look like
such great bargains because a portion of their price is tied to gas). You wou
request QFs to maximize their output. But credit problems have created an
PG&E and Edison have not been paying the QFs fully and promptly for so
creditors committee and threaten to push PG&E and Edison into bankruptc
shut down because they did not have money to pay for the gas.) ? Last wee
MidAmerican/CalEnergy to essentially sell its power to others even though
output to the purchasing utility. ? CalEnergy does so immediately, selling t
court ordered Reliant to continue to sell to the ISO, when requested, regard
paid fully and promptly for past deliveries to the ISO. Reliant announced it
somewhat of a contrast to the QF situation except that the circumstances go
probably different. The QF contracts pre-date the ISO and are with the utili
reference to providing power during emergencies. In fact, many QF contra
authority for the utility to cut takes during so-called "light load" periods. S
blackouts--again There was another Stage 3 emergency in California ? with
prompted everyone to wonder why this was happening in March. Among th
from summer-like temperatures ? Cutbacks in imports ? Loss of 1400 MW
Edison plant ? Loss of about 3100 MW from QF plants that were forced to

#	Summary	Sender	Date	ID
1	J **Mara**/NA/**Enron**@**Enron**, **Mike**	steven.kean@enron.com	2001030...	478
2	, Dow **Jones** & Company, **Inc**.) **California's**	suzanne_nimocks@mckinsey.com	2001032...	536
3	)=20 **California electricity prices** have shot back	miyung.buster@enron.com	2001062...	74
4	@aol.com, **James** = D=20	steven.kean@enron.com	2001042...	421
5	. ?????The **state's two** major utilities, **Pacific** Gas &	joseph.alamo@enron.com	2001042...	100
6	**electricity prices won**'t necessarily mean	jmunoz@mcnallytemple.com	2001043...	62
7	to the **ISO** report, **five** in-**state** power suppliers	miyung.buster@enron.com	2001032...	694
8	with the matter **told** Dow **Jones** Newswires late	miyung.buster@enron.com	2001032...	697
9	**billion** to the cost of **electricity** in the **state** since	jmunoz@mcnallytemple.com	2001052...	66
10	**billion** since **June** to high wholesale **prices** that the	jmunoz@mcnallytemple.com	2001041...	56
11	J Kean/NA/**Enron**@**Enron**, **Susan** J	steven.kean@enron.com	2001022...	487
12	to the **state's** investor-owned utilities, **Pacific** Gas	miyung.buster@enron.com	2001060...	72
13	& Company, **Inc**.) LOS ANGELES -(Dow **Jones**)-	miyung.buster@enron.com	2001032...	696
14	**billion** on power by **June** 2002.=20 But if	miyung.buster@enron.com	2001050...	63
15	**electricity prices** in **California**. The economists	jmunoz@mcnallytemple.com	2001052...	68
16	**prices** put **Pacific** Gas & **Electric** Co. and Southern	jmunoz@mcnallytemple.com	2001042...	57
17	**electricity** rate increase in **California** history,	kevinscott@onlinemailbox.net	2001052...	768
18	Davis says **California won**'t settle for $1 **billion** SF	angela.wilson@enron.com	2001071...	88
19	would have received **higher prices** under the bill	miyung.buster@enron.com	2001030...	464
20	, Fri, 7/6: **California** Changes Stance on **Refunds**;	miyung.buster@enron.com	2001070...	83
21	in **electricity price**= s,=20 while Hebert faulted	miyung.buster@enron.com	2001070...	82
22	Times, **October** 27 2000 **California** Adopts	ann.schmidt@enron.com	2000103...	379
23	.com, kgjohnson@bpa.**gov**,	steven.kean@enron.com	2001032...	871
24	for "01 and **John** has joined **Linda Robertson's**	steven.kean@enron.com	2001030...	461
25	L.; Novosel, **Sarah**; **Steffes**, **James** D. Subject: FW:	john.shelk@enron.com	2001103...	627
26	**higher electricity prices** in **California** and sales...	miyung.buster@enron.com	2001072...	92
27	of **electricity** were playing "games" to get **higher**	karen.denne@enron.com	2001111...	760
28	cc: **Linda Robertson**/NA/**Enron**@**ENRON** Subject:	steven.kean@enron.com	2001030...	465
29	@EES, **James** D **Steffes**/HOU/EES@EES,	steven.kean@enron.com	2000091...	821
30	= S,=20 **James** D **Steffes**/NA/**Enron**@**ENRON**,	steven.kean@enron.com	2001022...	483
31	, **Susan** J **Mara**/NA/**Enron**@**ENRON**, **Paul**	steven.kean@enron.com	2001022...	486
32	, **skean**@**enron**.com@**ENRON**, **Susan** J	steven.kean@enron.com	2001022...	489
33	.com>, **Jim Steffes**	steven.kean@enron.com	2001070...	81
34	**Jones** & Company, **Inc**.) **WASHINGTON** -- Angry	steven.kean@enron.com	2001030...	467
35	@**Enron**, **Susan** J **Mara**/NA/**Enron**@**ENRON**, **Joe**	steven.kean@enron.com	2001042...	418

Fig. 2. Screenshot of the list with the most-important documents.

The evaluation of clustering algorithms can be conducted with statistical measures as introduced in [32]. Those methods assess the quality of the clustering algorithms. However, the quality of the resulting clustering needs to be judged by human users. A reduced Enron dataset[4] is used to show the functionality of the developed module. This dataset consist of emails collected from Enron employees during the Enron scandal. Figure 3 shows the top-level clusters of this dataset. Important facts on the Enron scandal can be grasped at a glance without further knowledge on the background of this dataset. The system identifies Enron's Executive Vice President Steven J. *Kean* as one of the key players in the scandal in which *attorneys* and *New York* played a crucial role.

[4] Available at http://www.cs.cmu.edu/ enron/.

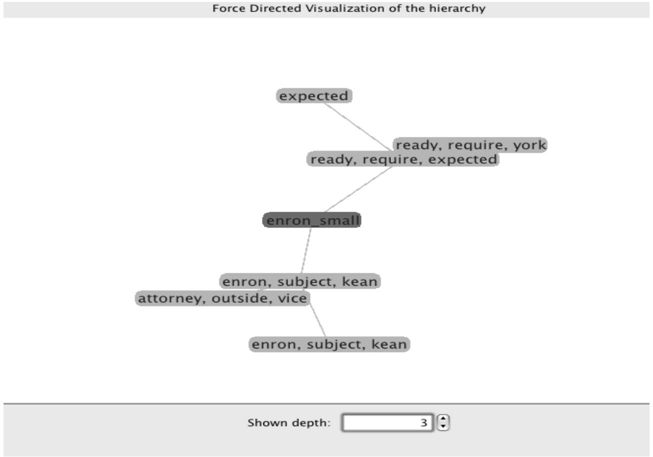


Fig. 3. Screenshot of the SNA-weighted top-level term clusters in the Enron dataset.

6 Discussion and Conclusion

A combination of information retrieval means and social network analysis techniques is introduced in this paper. The aim is to reveal discussed topics in social networks like email archives and their relationships among each other. Instead of relying only on information retrieval techniques the structure of the underlying social network is taken into account.

Foundations of information retrieval and social network analysis are described, techniques of both fields are combined to obtain taxonomies of the topics discussed in the communication of social networks. Instead of solely relying on methods of IR when determining concepts discussed in the communication archives, the structure of the underlying network is respected by factoring in SNA key measures.

Unsupervised clustering algorithms are used to scrutinize the content of communication in social networks. Classification of terms with those algorithms is a new approach to gain insights on communication networks. The resulting taxonomies of terms can be used to obtain an overview on the whole communication network at a glance. A temporal analysis module allows assessing the development of discussed topics over time. Finally, the design and details of the implementation are presented.

The aim of this paper was to provide users of social network analysis packages like Condor with an automated method to reveal topics discussed in analyzed networks and their hidden relations among each other. First steps were made in this paper to allow users to gain an impression on the nature of discussions in analyzed networks. This work can only serve as a step in the right direction of automatically revealing discussed topics in social networks. Supporting users with enhanced automated or semi-automated methods to analyze the content of social networks is crucial with more and more data available on social networks.

References

1. Boyd, D., Ellison, N.B.: Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication **13**(1-2) (November 2007)
2. Gloor, P.A., Cooper, S.M.: Coolhunting: chasing down the next big thing. Amacom, New York, NY, USA (2007)
3. Borgatti, S.P., Foster, P.C.: The network paradigm in organizational research: A review and typology. Journal of Management **29**(6) (December 2003) 991–1013
4. Korfiatis, N., Sicilia, M.A., Hess, C., Stein, K., Schlieder, C.: VI. In: Social Network Models for Enhancing Reference Based Search Engine Rankings. Idea Group Reference (2007) 109–133
5. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (2000)
6. Cai, L., Hofmann, T.: Hierarchical document categorization with support vector machines. In: CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management, New York, NY, USA, ACM (2004) 78–87
7. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. Journal of the American Society of Information Science **41**(6) (1990) 391–407
8. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM **18**(11) (1975) 613–620
9. Scott, J.P.: Social Network Analysis: A Handbook. SAGE Publications (January 2000)
10. Freeman, L.C.: Social Network Analysis: Definition and History. Encyclopedia of Psychology **6** (2000) 350–351
11. Krackhardt, D., Brass, D. In: Intra-Organizational Networks: The Micro Side. Sage Publications (1994) 209–230
12. Klovdahl, A.S.: Social network research and human subjects protection: Towards more effective infectious disease control. Social Networks **27**(2) (2005) 119–137
13. Gloor, P.A., Zhao, Y.: Analyzing actors and their discussion topics by semantic social network analysis. In: IV '06: Proceedings of the conference on Information Visualization, Washington, DC, USA, IEEE Computer Society (2006) 130–135
14. Carley, K.M.: Dynamic network analysis. In Breiger, K.C.R., Pattison, P., eds.: Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers. Committee on Human Factors, National Research Council (2003) 361–370
15. Gloor, P.A.: Capturing team dynamics through temporal social surfaces. In: IV '05: Proceedings of the Ninth International Conference on Information Visualisation, Washington, DC, USA, IEEE Computer Society (2005) 939–944
16. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. **31**(3) (1999) 264–323
17. Kowalski, G., Maybury, M.T.: Information Storage and Retrieval Systems: Theory and Implementation. Kluwer Academic Publishers, Norwell, MA, USA (2000)
18. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of semantic distance. Computational Linguistics **32**(1) (2006) 13–47

19. Li, Y., Bandar, Z., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on Knowledge and Data Engineering **15**(4) (2003) 871–882
20. Schonhofen, P.: Identifying document topics using the wikipedia category network. In: WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, Washington, DC, USA, IEEE Computer Society (2006) 456–462
21. Nastase, V., Strube, M.: Decoding wikipedia categories for knowledge acquisition. In Fox, D., Gomes, C.P., eds.: AAAI, AAAI Press (2008) 1219–1224
22. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. Systems, Man and Cybernetics, IEEE Transactions on **19**(1) (Jan/Feb 1989) 17–30
23. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: International Joint Conference for Artificial Intelligence (IJCAI-95). (1995) 448–453
24. Lin, D.: An information-theoretic definition of similarity. In: ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (1998) 296–304
25. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (04 2008)
26. Sneath, P.H.A., Sokal, R.R.: Numerical taxonomy: the principles and practice of numerical classification. Freeman, San Francisco, USA (1973)
27. Cutting, D.R., Karger, D.R., Pedersen, J.O., Tukey, J.W.: Scatter/gather: a cluster-based approach to browsing large document collections. In: SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (1992) 318–329
28. Gloor, P.A., Zhao, Y.: Tecflow - a temporal communication flow visualizer for social networks analysis. In: CSCW'04 Workshop on Social Networks, ACM (2004)
29. Gloor, P.A., Oster, D., Putzke, J., Fischbach, K., Schoder, D., Ara, K., Kim, T., Laubacher, R., Mohan, A., Olguin Olguin, D., Pentland, A., Waber, B.N.: Studying microscopic peer-to-peer communication patterns. In: Proceedings AMCIS Americas Conference on Information Systems. (2007)
30. Fruchterman, T.M.J., Reingold, E.M.: Graph drawing by force-directed placement. Softw. Pract. Exper. **21**(11) (1991) 1129–1164
31. Eades, P.A.: A heuristic for graph drawing. In: Congressus Numerantium. Volume 42. (1984) 149–160
32. Stein, B., zu Eissen, S.M., Wißbrock, F.: On cluster validity and the information need of users. In: Proceedings of the Artificial Intelligence and Applications Conference. (2003)