Mining the Temporal Dimension of the Information Propagation

Michele Berlingerio[1], Michele Coscia[2], and Fosca Giannotti[3]

[1] IMT-Lucca, Lucca, Italy
[2] Dipartimento di Informatica, Pisa, Italy
{name.surname}@isti.cnr.it
[3] ISTI-CNR, Pisa, Italy

Abstract. In the last decade, Social Network Analysis has been a field in which the effort devoted from several researchers in the Data Mining area has increased very fast. Among the possible related topics, the study of the information propagation in a network attracted the interest of many researchers, also from the industrial world. However, only a few answers to the questions "How does the information propagates over a network, why and how fast?" have been discovered so far. On the other hand, these answers are of large interest, since they help in the tasks of finding experts in a network, assessing viral marketing strategies, identifying fast or slow paths of the information inside a collaborative network. In this paper we study the problem of finding frequent patterns in a network with the help of two different techniques: TAS (Temporally Annotated Sequences) mining, aimed at extracting sequential patterns where each transition between two events is annotated with a typical transition time that emerges from input data, and Graph Mining, which is helpful for locally analyzing the nodes of the networks with their properties. Finally we show preliminary results done in the direction of mining the information propagation over a network, performed on two well known email datasets, that show the power of the combination of these two approaches.

1 Introduction

In the last decade, the interest in Social Network Analysis topics from researchers in the Data Mining area has increased very fast. Much effort has been devoted, for example, in the Community Discovery, Leader Detection and Network Evolution problems [7,18,4,17]. Another topic that has attracted much interest recently is how the information propagates over a network [10,1,14,13]. This problem has been studied from several points of view: statistics, modeling, mining are few of the approaches that have been applied so far in this direction. However, only a few answers to the questions "How does the information propagates over a network, why and how fast?" have been discovered so far. On the other hand, these answers are of large interest, since they help in the tasks of finding experts in a network, assessing viral marketing strategies, identifying fast or slow paths

N. Adams et al. (Eds.): IDA 2009, LNCS 5772, pp. 237–248, 2009.

of the information inside a collaborative network, and so on. In this paper we study the problem of finding frequent patterns in a network focusing in two aspects:

- The temporal dimension intrinsically contained in the flow of information: why certain topics are spread faster than others? What is the distribution of the temporal intervals among the “hops” that the information passes through?
- The causes of the information propagation: why certain discussions are passed over while others stop in two hops? What are the characteristics of the nodes that pass the information?

As one can notice, the two dimensions of our focus are orthogonal to each other: certain nodes with certain characteristics may let a particular kind of information spread faster or slower than other nodes, or compared to information with other characteristics. The combination of the two aspects finds several possible application in real life. Among all of them, we believe that Viral Marketing can be powerfully enhanced by such kind of analysis. Companies willing to advertise a new product in their network of users may discover that giving a certain kind of information or special offers to a particular set of selected nodes may result in a cheaper or more effective advertisement campaign.

In this paper, we study the above problem on the well known Enron email dataset [12], and the 20 Newsgroups dataset [11,16], and with the help of two different techniques: TAS (Temporally Annotated Sequences) mining, which is a paradigm aimed at extracting sequential patterns where each transition between two events is annotated with a typical transition time that emerges from input data, and Graph Mining, which is helpful for locally analyzing the nodes of the networks with their properties.

The contribution of this paper can be summarized as follows: we show how to extract useful information from a network in order to mine the information propagation, in the format of a graph where nodes are users and edges are words used as email subjects, and a set of timestamped sequences of emails grouped by threads; we show how to apply the two techniques above to a real-life dataset; we present the preliminary results obtained by applying the two algorithms on graphs extracted from the datasets, and on the sequences of exchanged email, showing a general methodology that can be applied in any sort of network where an exchange of information is present.

The rest of the paper is organized as follows: section 2 presents some work related to our problem; section 3 defines what is the problem under investigation and which kind of data we want to analyze; section 4 shows the preliminary results obtained during our analysis of the datasets; section 5 briefly summarizes the results of our work and some possible future work.

2 Related Work

During the last years, several approaches have been proposed addressing the problem of analyzing how the information propagates in a network.

In [10], the authors summarize three papers focusing on finding communities and analyzing the small world phenomenon by means of statistical approaches. In [1], the authors describe general categories of information epidemics and create a tool to infer and visualize the paths specific infections take through the network by means of statistical tools and Support Vector Machines. In [6], the model of timestamped graph and digraph are introduced in order to study the influence in a network. In [14] the problem of Viral Marketing is analyzed with several different statistical approaches.

Among several other possible works, we believe that [13,15] are the closest to our work: they focus on the temporal behavior in the network, and in the characteristic of the users in the network.

However, to the best of our knowledge, this is the first time that TAS mining and Graph Mining are used in conjunction in order to tackle the problem of finding frequent patterns of information propagation together with their causes in a network.

3 Problem Definition

We are given a dataset $\mathcal{D}$ of activities in a network, from which we can extract both a network of users $\mathcal{U}$ as a graph $\mathcal{G}$ and a flow of any kind of information (emails, documents, comments, instant messages, etc.) as a set of timestamped sequences $\mathcal{S}$. Examples of such datasets can be a set of emails exchanged among people, the logs of an instant messaging service, the logs of a social networking system, the content of a social bookmarking site, and so on. In this dataset we are interested in finding frequent patterns of information propagation, and we want to let the causes of such patterns emerge from the data. This can be done by applying a framework for extracting temporally annotated sequences, as shown in section 4, which allow to find such causes modeled as itemsets. We then want to compare these rich patterns with the local patterns found in the graph $\mathcal{G}$, to see how the characteristics of the nodes interact both with the information spread and with the interactions of the nodes with their local communities in the network.

We assume $\mathcal{D}$ to contain at least the information about:

- a set of users with their characteristics (such as: gender, country, age, typical discussed topics, degree, betweenness and closeness centrality computed over the network, and so on)
- a timestamped set of sequences of actions performed by the above users that involve the propagation of a certain kind of information (such as: exchange of emails, posts in a forum, instant messages, comments in a blog, and so on)

From the first, we can build several kinds of graphs that can be analyzed with classical graph mining techniques. In order to mine and analyze the local communities of the nodes with the focus on the spread of information, we want to build such graphs on the basis of the information exchanged among the nodes.

As an example, the nodes of the graph can be the users of a mailserver, while there is an edge between two nodes if the nodes exchanged an email. The edge can be then labeled with the typical words used in the communications, that may be also grouped semantically or by statistical properties. Depending on the characteristics of the users and the way we consider them connected among each other, we are able to perform Social Network Analysis of the original network from several different points of view. For example, we may want to use as vertex labels the gender, the country and the age if we are analyzing a so called web social network, while we may want to use structural properties such as the degree, the closeness centrality, the betweenness or the clustering coefficient, if we are analyzing a network of a company. Each different combination of properties would result in a different kind of analysis.

From the second, we can derive a set of timestamped sequences to use as an input of the TAS mining paradigm (see Section 4), in order to be able to extract sequences of itemsets (i.e. characteristics of the users) that are found frequent in the data, together with frequent temporal annotations for them.

The entire analysis will be an interactive and iterative loop of the following steps:

1. Building a graph $\mathcal{G}$ of users in $\mathcal{U}$, connected by edges representing typical words or topics discussed by or among them
2. Assigning labels to the users in $\mathcal{U}$ according to their semantical (such as age, gender, newsgroup of major activity, preferred topic, etc.) and statistical (computed in $\mathcal{G}$, such as betweenness, closeness centrality, etc.) characteristics, collecting them in a set $\mathcal{L}$
3. Assigning labels to the edges in $\mathcal{G}$ according to their semantical (such as semantical cluster, etc.) or statistical (such as frequency of the stemmed word or topic in the subjects, etc.) characteristics, collecting them in a set $\mathcal{W}$
4. Extracting the flows of information in $\mathcal{D}$, grouped by any property to use as transaction identifier (thread, email subject, conversation ID, ..), and building a set of temporally annotated sequences $\mathcal{S}$, containing both the information on the users involved in each flow (represented as itemsets of labels in $\mathcal{L}$), and the temporal information about the flow (usually found as timestamps in seconds since the Epoch)
5. Extracting frequent Temporally Annotated Sequences $\mathcal{T}$ from $\mathcal{S}$, representing the frequent flows of information, and containing both the temporal dimension of the patterns, and the characteristics of the users involved
6. Extracting frequent subgraphs from $\mathcal{G}$ with the help of classical Graph Mining, that represent the local communities of nodes together with their characteristics and typical words or topics used
7. Analyzing the results produced in 4 in order to find frequent items (users' caracteristics) associated with typical fast or slow transition times, then analyze the patterns produced in 6 in order to find patterns containing nodes with the same characteristics as labels: these patterns will tell if the users with these characteristics are the best ones in spreading fast the type of information described by the graph patterns

Steps 1, 2, and 3, are clearly crucial and may vary the analysis that will be performed. By setting different labelings for the edges in $\mathcal{G}$ and including or excluding different characteristics as vertex labels in 6, the analyst may drive the search for frequent patterns in different directions. Please note that, due to the techniques available nowadays to the best of our knowledge, while the TAS mining framework allows for the use of itemsets, which represent a set of characteristics, there appears not to exist a graph miner able to handle more than one label per edge. Hence, step 3 basically implies that we have to produce different input graphs for every kind of analysis we want to perform.

4 Case Study

4.1 Dataset

We used for our experiments two e-mail datasets. The first one is the Enron email dataset[12]. This dataset contains 619,446 email messages complete with senders, recipients, cc, bcc, and text sent and received from 158 Enron's employees. This dataset is characterized by an exceptional wealth of information, and it allows to track flows of communication, together with their associated subjects and the complete data regarding the exchange of information. We took from the entire dataset the "from", "to", "cc", "bcc", "subject" and "date" fields in each email in the "sent" folder of every employee. We took only the emails that were sent to other Enron employees, removing the outgoing emails. We also performed basic cleaning by removing emails with empty subjects, noise, and so on. After the cleaning stage, the number of remaining emails was about 12,000. We refer to it as the "Enron" dataset.

The second dataset consists of Usenet articles collected from 20 different newsgroups about general discussions on politics and religion, technical discussions on computers and hardware, general discussions on hobbies and sports, general discussion on sciences, and a newsgroup for items on sale, and was first used in [11,16]. Over a period of time, 1000 articles were taken from each of the newsgroups, which makes an overall number of 20,000 documents in this collection. Except for a small fraction of the articles, each document belongs to exactly one newsgroup. We took from each sent email the "from", "to" and "date" field. After a cleaning stage, the number of remaining emails was about 18,000. We refer to it as the "Newsgroup" dataset.

4.2 Tools

For our analysis, we used the MiSTA software [9,8], which extracts frequent Temporally Annotated Sequences from a dataset of timestamped sequences, and that has been successfully applied in several contexts [3,2]; we also used a single graph miner in order to find frequent subgraphs of a large graph, implementing a Minimum Image Support function as described in [5].

All the experiments were conducted on a machine equipped with 4 processors at 3.4GHz, 8GB of RAM, running the Ubuntu 8.04 Server Edition, and took from seconds to minutes for the TAS mining, and from minutes to hours for the graph mining.

4.3 Steps of Analysis

We then followed the steps described in section 3 in order to perform our analysis. In the following steps, the subscripts E and N indicate whether the sets refer to the Enron or Newsgroup datasets, respectively.

As step 1, we built the graph $\mathcal{G}_{\mathcal{E}}$ for the Enron dataset by taking the users as nodes and connecting two nodes with edges representing the subjects of emails exchanged between them. For the Newsgroup dataset, we built $\mathcal{G}_{\mathcal{N}}$ by taking the users as nodes and connecting two nodes with edges representing the subjects for which both users posted a message to the newsgroups.

As step 2, we labeled the users $\mathcal{U}_{\mathcal{E}}$ and $\mathcal{U}_{\mathcal{N}}$ following five different possible labeling, according to their structural characteristics in the graphs $\mathcal{G}_{\mathcal{E}}$ and $\mathcal{G}_{\mathcal{N}}$: the degree (the number of ties to other nodes in the network, referred as "DEG"), the closeness centrality (i.e., the inverse of the distance in number of edges of the node from all other nodes in the network, referred as "CL"), the betweenness centrality (i.e., the number of geodesic paths that pass through the node, referred as "BET") and two different clustering annotations (the first, referred as "CC1", is the triadic closure ratio, while the second, referred as "CC2", is a modified version that privileges the 2-neighborhood clustering). Table 1 shows the labeling according to the real values of these variables. For users in $\mathcal{U}_{\mathcal{N}}$ we also performed a labeling according to the newsgroup in which the user was most active, assigning thus 20 possible labels for each node.

As step 3, the edges in $\mathcal{G}_{\mathcal{E}}$ and $\mathcal{G}_{\mathcal{N}}$ have been assigned a label according to various criteria. Both for the Enron and the Newsgroup datasets, the most frequent words in the subjects were manually clustered by their semantic in 5 different clusters per dataset. Each edge was then labeled with the most frequent cluster

Table 1. The labels assigned to the users in the datasets

Enron					
Label	Degree	Closeness	Betweenness	CC1	CC2
1	$[0-5]$	$[0-0.21[$	$[0-0.0015[$	0	0
2	$[6-15]$	$[0.21-0.2329[$	$[0.0015-0.0046[$	$]0-0.2[$	$]0-35e^{-6}[$
3	$[16-33]$	$[0.2329-0.2513[$	$[0.0046-0.013[$	$[0.2-0.34[$	$[35e^{-6}-14e^{-5}[$
4	$[34-75]$	$[0.2513-0.267[$	$[0.013-0.034[$	$[0.34-0.67[$	$[14e^{-5}-61e^{-5}[$
5	$[76-+\infty[$	$[0.267-1]$	$[0.034-1]$	$[0.67-1]$	$[61e^{-5}-1[$

Newsgroup					
Label	Degree	Closeness	Betweenness	CC1	CC2
1	$[0-15]$	0	0	0	0
2	$[16-39]$	$]0-0.12[$	$]0-0.0002[$	$]0-0.42[$	$]0-0.00015[$
3	$[40-84]$	$[0.12-0.145[$	$[0.0002-0.001[$	$[0.42-0.61[$	$[0.00015-0.00085[$
4	$[85-154]$	$[0.0002-0.001[$	$[0.001-0.002[$	$[0.61-1[$	$[0.00085-0.005[$
5	$[155-+\infty[$	$[0.1632-1]$	$[0.002-1]$	1	$[0.005-1[$

Table 2. The dataset statistics

Graph	n	e	$\bar{k}$	#Components	GiantComponent	$\bar{C}$	ℓ	Diameter
Enron S	3731	9543	5.11	30	98.01%	0.17	4.52199	15
Newsgroup S	1457	12560	17.24	151	64.51%	0.78	4.02730	11
Newsgroup F	3923	31632	16.12	249	82.41%	0.73	4.42142	17

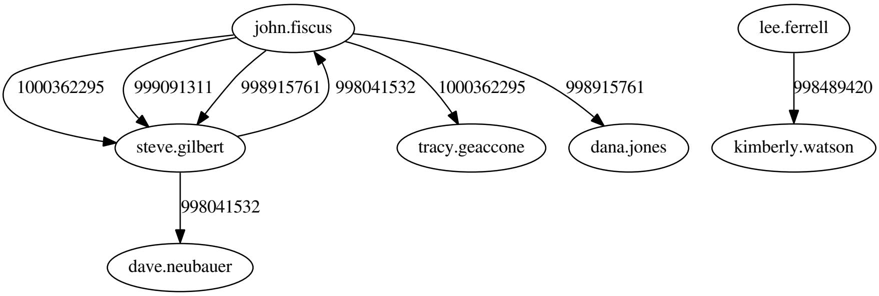


Fig. 1. Example of mail flow for the subject "2002 capital plan"

among its words (ignoring the words not belonging to any cluster). The edges corresponding to subjects for which none of the contained words was frequent or was not belonging to any of the clusters were removed from the graphs. We refer to the graphs created in this way as "Enron S" and "Newsgroup S". For the Newsgroup dataset we performed also a different labeling: all the words were divided in three frequency classes and the edges were then labeled accordingly. We refer to this graph as "Newsgroup F". Finally, in each graph, multiple edges between two nodes have been collapsed into a single edge labeled with its more frequent label. Table 2 shows some statistical properties of the graphs generated, in which: n is the number of nodes, e the number of edges, $\bar{k}$ the average degree, *#Components* the number of components, *GiantComponent* the size of the largest component of the graph (percentage of the total number of nodes), $\bar{C}$ the average clustering coefficient of the graph (between 0 and 1), ℓ the average length of the shortest paths in the graph and *Diameter* the length of the longest shortest path in the graph.

For the step 4, in order to build our $\mathcal{S}_{\mathcal{E}}$ and $\mathcal{S}_{\mathcal{N}}$ for the TAS mining paradigm, we grouped all the emails by subject, keeping the timestamp given by the mailserver to every email. Figure 1 is a graphical representation of the flow of emails in Enron with initial subject "2002 capital plan". In order to give this in input to the software, we processed each of these flow by splitting it in all the possible sequences of emails passed from an user to the others, following the natural temporal ordering. This last step was not necessary for Newsgroup, as the emails were sent only to one recipient, namely the newsgroup. The complete set of these timestamped sequences constituted then our $\mathcal{S}_{\mathcal{E}}$ and $\mathcal{S}_{\mathcal{N}}$.

Steps 5 and 6 produced the results in the following paragraph.

4.4 Results

Although the focus in this paper is only to show the power of the combination of the two techniques we used in our analysis, we can make some interpretation of some of the resulting patterns, that clearly show the differences between the two datasets.

Graph Mining
The Enron graph represents interactions in the working environment of a company, from which we can infer particular considerations regarding possible stages of the workflow followed by the employees. Contacts between employees are direct (not thus as in the newsgroup case), and are very often one-to-many (i.e. there are many recipients in cc in an email).

The first pattern extracted, Figure 2a, represents an exchange of emails. The labels on the nodes represent the level of Clustering Coefficient 2, i.e. the tendency of an employee to create a working group around him or her. It can be noticed that employees with high CC2 have a frequent exchange of emails among each other with several subjects. At a certain moment, one of these high CC2 employees has a contact with a lower CC2 employee (a node outside the central part of the graph being maybe a specific member of a work group) with a different subject (label 4 vs other labels in the pattern). This pattern may represent the mechanism by which members acting as "bridges" between groups detect, with a mutual exchange of knowledge, who can solve a problem.

Figure 2b, where the label "<5" means any label lower than 5, is a generalization of Figure 2a. We found several patterns that follows this generalization, thus we consider the phenomenon described above quite interesting in this case study.

Another interesting pattern in the Enron graph is represented in Figure 2c, where labels are assigned to nodes according to the first definition of clustering coefficient (CC1). One can see that nodes with a low clustering coefficient (discovered to be synonymous of high degree, in our case study) tend to behave in contrast to the value of such a coefficient, as the nodes are found in a frequent clique. This happens because these nodes have a very high degree (i.e. they represent managers and directors), and they are all connected by edges labeled with 1, which represents subjects regarding high-level decisions in the company. In other words, cliques among managers are frequent only if they are speaking about high-level topics.

We conclude the discussion on frequent graph pattern noting that, among the results obtained in the Newsgroup dataset by using as labels for the edges the

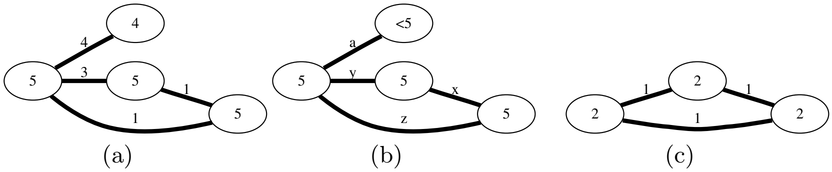


Fig. 2. Subgraphs found in Enron dataset

frequencies of the words composing the subjects, and the value of the CC2 as labels for the nodes, users with a specific CC2 tend to speak about a specific class of subjects with other users having the same CC2, while they speak about other subjects when talking to people with a different CC2. Examples of this behavior are patterns in Figure 3a and 3b.

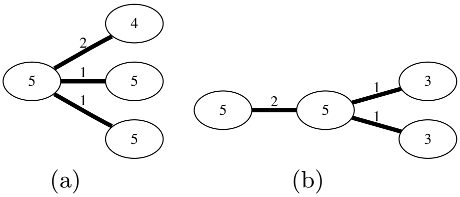


Fig. 3. Subgraphs found in Newsgroup dataset

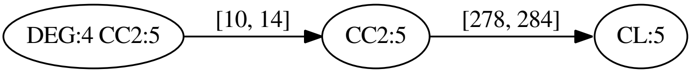


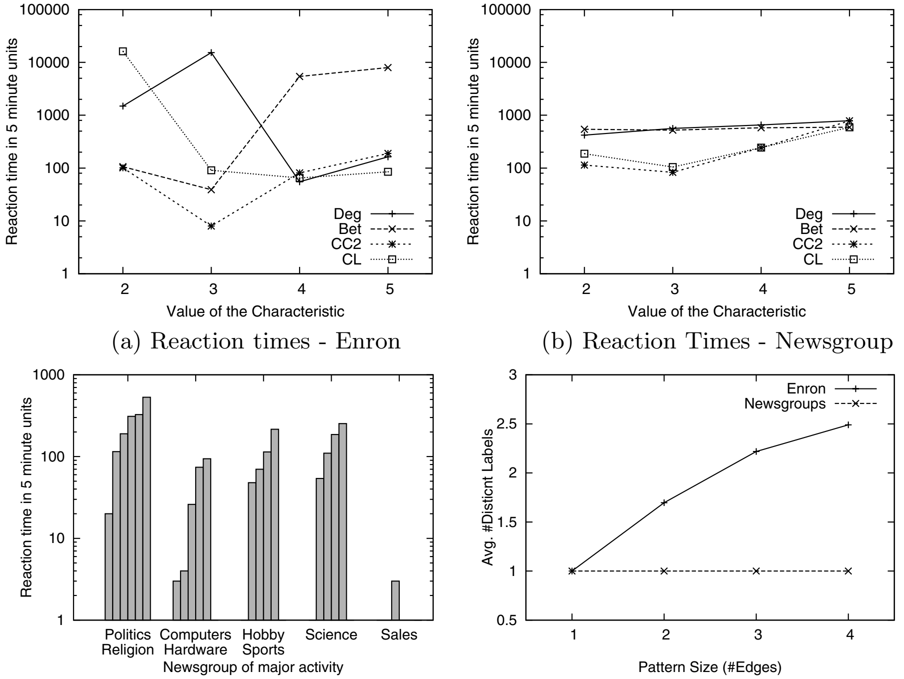
Fig. 4. An example of TAS found

TAS

We now present some considerations derived from the analysis of the most frequent temporal sequences extracted from the two datasets. Figure 4 is a graphical representation of a possible extracted pattern, saying that a user with DEG=4 and CC2=5 replied to an email 10 to 14 time units (5 minutes each) afterwards, to a recipient with CC2=5, which replied to the same email 278 to 284 time units afterwards, to a user with CL=5.

Consider graphs in Figure 5. The 5a and 5b graphs were generated by analyzing the average response time of the most frequent sequences (i.e. the most representative) according to different characteristics (Degree, Closeness, Betweenness and CC2) of the sender. First, we can notice the difference of reaction times, found to be higher in average in the Enron dataset. This can be explained by considering the different nature of the exchange of knowledge in a working environment: there are not (frequent) immediate answers, since, after an email, usually there are several steps of gathering of information, meetings and brainstorming, thus enlarging the time needed for providing a response. On the opposite side, users within a social community usually only need to read all the messages before answering, leading to usual short response time.

Regarding the Enron dataset, Figure 5a reveals an important piece of information regarding the response time of the employees with an high degree of betweenness centrality: having such an high centrality for an employee means that many shortest paths in internal communications pass through that employee. TAS mining revealed that this tends to result in much higher response times, due to the additional working burden that employees of this type have to face. The knowledge that can be extracted from this analysis is to avoid, if



(a) Reaction times - Enron

(b) Reaction Times - Newsgroup

(c) Reaction times - Newsgroup

(d) Graph Patterns Heterogeneity

Fig. 5. Quantitative Analysis of the Results

possible, the structural hubs when there is the need to speed up a communication. This correlation between reaction times and betweenness was not found frequent in the Newsgroup dataset.

Regarding the Newsgroup dataset, Figure 5b shows another difference of behavior from the Enron dataset: the higher regularity of growth of the reaction time for users with higher degree. The degree grows as the user follows many different discussions and, especially, when these discussions involve an increasing number of users. The results showed that the typical response times go up because these discussions are probably the most controversial and interesting ones, and in order to follow them, much time and attention have to be spent. These considerations are not necessarily true in a business context: an employee with a high degree (many different contacts) often answers very quickly.

Another consideration can be done w.r.t the "newsgroup" labeling. Consider Figure 5(c). In the x axis we have the newsgroup of major activity of the users (i.e., a possible value of the "newsgroup" label for the nodes), clustered by main topics (x labels), while in the y the reaction times as found in the frequent TAS. Each of the 20 bars represents one particular newsgroup. As we can notice, there are differences in the reaction times according to the main topic of the newsgroups. While politics and religion seem to be general "relaxed" discussion topics, technical discussions in computers and hardware find more reactive answers. The most reactive is the newsgroup where people put items for sale: the first offer is generally set after 5-10 minutes, as we can see from the figure.

Based on the above consideration, we can give a “draft” of what could be done in order to perform step 7 of the general approach described in section 3: once found that the “Sale” label could be a characteristic related to the speed of the users, it is possible to go back on the results of the graph mining and see if that label was found frequent, possibly in the center of a large subgraph pattern where other nodes have different labels. If the frequency of this pattern is found high, one can argue that passing the information to “Sale” nodes would result in a faster and effective spread of information. In this case study, the meaning of the “Sale” label does not really suggest anything special, but the focus here is to give an idea of the potentialities of the general approach followed.

Finally, consider Figure 5d that shows a direct comparison between the two datasets. It shows the degree of heterogeneity of the communication (i.e. the number of different semantic labels associated with the edges of the frequent patterns) compared with the volume of communication (number of edges in the pattern). From the graph it can be inferred that the business environment shows a greater heterogeneity in the communications: while in the Enron dataset employee tend to speak about different topics with their neighbors, in the Newsgroup dataset close users speak about the same topics. This seems quite easy to explain: employees usually manage more than one different situation, while users in newsgroups tend to be clustered by newsgroup, and hence by topics.

5 Conclusions and Future Work

We have shown a general methodology to mine the information propagation in a network where users exchange information. We have described how to extract useful information from such a network in order to be able to use a combination of two powerful techniques, namely TAS mining and graph mining, in order to find frequent patterns of propagation of information that involve also the possible causes of this propagation. We have shown how this combination can help in finding frequent temporal behaviors in the network together with the characteristics of the users, and what are the roles of these users in the network. We have presented preliminary results of a case study on real-life datasets and we have provided a possible interpretation of some of them.

These first results are encouraging and open the way for a powerful methodology that can help applications such as Viral Marketing.

In the future we plan to extend this analysis to other datasets, where we will be able to use more characteristics of the words and different characteristics of the users, such as the country, the gender, the age and so on.

References

1. Li Zhang, L.A., Adamic, R.M., Lukose, E.A.: Implicit structure and the dynamics of blogspace. Communications of the ACM: CACMa publ. of the Association for Computing Machinery 47(12), 35–39

2. Berlingerio, M., Bonchi, F., Giannotti, F., Turini, F.: Mining clinical data with a temporal dimension: a case study. In: Proc. of The 1st Intern.Conf. on Bioinf. and Biomed. (2007)
3. Berlingerio, M., Bonchi, F., Giannotti, F., Turini, F.: Time-annotated sequences for medical data mining. In: Proc. of The Intern. Workshop of Data Min. in Medicine (2007)
4. Borgwardt, K.M., Kriegel, H.-P., Wackersreuther, P.: Pattern mining in frequent dynamic subgraphs. In: IEEE International Conference on Data Mining, pp. 818–822 (2006)
5. Bringmann, B., Nijssen, S.: What is frequent in a single graph? In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 858–863. Springer, Heidelberg (2008)
6. Cheng, E., Grossman, J.W., Lipman, M.J.: Time-stamped graphs and their associated influence digraphs. Discrete Appl. Math. 128(2-3), 317–335 (2003)
7. Desikan, P., Srivastava, J.: Mining temporally changing web usage graphs. In: Mobasher, B., Nasraoui, O., Liu, B., Masand, B. (eds.) WebKDD 2004. LNCS (LNAI), vol. 3932, pp. 1–17. Springer, Heidelberg (2006)
8. Giannotti, F., Nanni, M., Pedreschi, D.: Efficient mining of temporally annotated sequences. In: Proc. of the 6th SIAM Intern. Conf. on Data Min. (2006)
9. Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F.: Mining sequences with temporal annotations. In: Proc. of the 2006 ACM Symp. on Applied Comp. (SAC), pp. 593–597 (2006)
10. Huberman, B.A., Adamic, L.A.: Information dynamics in the networked world (October 2003)
11. Joachims, T.: A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In: Fisher, D.H. (ed.) Proceedings of ICML 1997, 14th International Conference on Machine Learning, Nashville, US, pp. 143–151. Morgan Kaufmann Publishers, San Francisco (1997)
12. Klimt, B., Yang, Y.: The enron corpus: A new dataset for email classification research. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) ECML 2004. LNCS (LNAI), vol. 3201, pp. 217–226. Springer, Heidelberg (2004)
13. Kossinets, G., Kleinberg, J., Watts, D.: The structure of information pathways in a social communication network (June 2008)
14. Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing (September 2005)
15. Liben-Nowell, D., Kleinberg, J.: Tracing information flow on a global scale using Internet chain-letter data. Proceedings of the National Academy of Sciences 105(12), 4633–4638 (2008)
16. Mitchell, T.: Machine Learning. McGraw-Hill Education (ISE Editions) (October 1997)
17. Sun, J., Faloutsos, C., Papadimitriou, S., Yu, P.S.: Graphscope: parameter-free mining of large time-evolving graphs. In: KDD 2007: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 687–696. ACM, New York (2007)
18. Tantipathananandh, C., Berger-Wolf, T., Kempe, D.: A framework for community identification in dynamic social networks. In: KDD 2007: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 717–726. ACM Press, New York (2007)