Extracting Semantic Knowledge from Twitter

Peter Teufl, Stefan Kraxberger

Extracting Semantic Knowledge from Twitter

Peter Teufl and Stefan Kraxberger
peter.teufl@iaik.tugraz.at, stefan.kraxberger@iaik.tugraz.at

IAIK, Graz University of Technology
Inffeldgasse 16a, 8010 Graz, Austria

Abstract. Twitter is the second largest social network after Facebook and currently 140 millions Tweets are posted on average each day. Tweets are messages with a maximum number of 140 characters and cover all imaginable stories ranging from simple activity updates over news coverage to opinions on arbitrary topics. In this work we argue that Twitter is a valuable data source for e-Participation related projects and describe other domains were Twitter has already been used. We then focus on our own semantic-analysis framework based on our previously introduced *Semantic Patterns* concept. In order to highlight the benefits of semantic knowledge extraction for Twitter related e-Participation projects, we apply the presented technique to Tweets covering the protests in Egypt starting at January 25^{th} and resulting in the ousting of Hosni Mubarak on February 11^{th} 2011. Based on these results and the lessons learned from previous knowledge extraction tasks, we identify key requirements for extracting semantic knowledge from Twitter.

Key words: Semantic Patterns, Twitter Mining, e-Participation, Semantic Analysis, Trend Analysis, Semantic Search, Machine Learning, Social Network Analysis

1 Introduction

A blog post from Twitter[1] reveals numbers that give us an impression of this social network that turned five years old in March 2011. Twitter states that at each day during the month before the blog entry an average of 140 million Tweets were posted and that 460.000 user accounts were added daily. While Twitter does not mention the current number of users, their latest statistics were released in June 2010 and stated that there were 190 million[2] users at this time. When comparing this with the 460.000 user accounts added per day, we can assume that Twitter has reached more than 200 million total users. This makes it the second largest social network after Facebook, which had reached the 500 million mark in July 2010.

The messages posted on Twitter are called *Tweets* and are comprised of maximum 140 characters. This is roughly similar to the 160 character limitation

[1] http://blog.twitter.com/2011/03/happy-birthday-twitter.html
[2] http://techcrunch.com/2010/06/08/twitter-190-million-users/

of the well established text messages (SMS) sent from our cell phones. Although, this limit seems to be rather short, it comes with a significant advantage – a user who posts a Tweet must carefully choose the terms and thereby compress the original information. This compression simplifies the manual and automated analysis of Tweets.

There are some basic concepts that are important for understanding how information is conveyed via the posted Tweets. *Followers* are persons that are interested in the Tweets of a specific user, whereas *friends* are other persons followed by a given user. *Retweeting* is the process of forwarding interesting Tweets to one's followers. Another important concept is the employment of *Hashtags*, which are arbitrary terms chosen by the users and preceded with a #. They are intended for the simple categorization of Tweets and allow the real-time monitoring of specific topics. Current examples are *"#Libya"*, *"#Egypt"* or *"#Syria"*. Since these hashtags are chosen by the community, they represent a self organizing process that evolves according to principles described by Halpin et al. [6]. In general, the information conveyed by Tweets covers all aspects of our society ranging from simple daily activities, over news coverage to discussions and opinions about arbitrary topics. Due to the facts that most of these Tweets are publicly available, that there is a huge user base and that all information must be compressed to 140 characters, Twitter represents a valuable resource for knowledge-mining. Twitter has already been called *The SMS of the internet*[3], but one could even go further and describe it as *The Online Presence of our Society*.

During the last three years we have focused on the development of a framework for the automated extraction of semantic knowledge. This framework is based on the a new concept called *Semantic Patterns* that we have already successfully deployed in a broad area of domains. Here we apply the framework to data extracted from Twitter.

The remainder of this work is organized as follows: In the next section we cover various Twitter related research projects that highlight the wide range of possible applications. We then give an introduction to the employed *Semantic Patterns* concept and present its key advantages. In the subsequent section, we address how Tweets can be extracted from Twitter, and finally we demonstrate the technique by analyzing the evolution of Tweets relevant to the Egyptian revolution. Due to our broad application of this technique, we are able to present the learned lessons which lead us to several key requirements that are also valid for e-Participation projects.

2 Related ideas and e-Participation use cases

Due to the abundance of data covering a wide range of topics, Twitter is a wealth for knowledge mining. The most obvious source for information is the text (including the hashtags) contained within the Tweets. For the analysis of

[3] http://www.business-standard.com/india/news/swine-flu%5Cs-tweet-tweet-causes-online-flutter/356604/

this information well-know processes from NLP, machine learning and statistics play an important role. However, in addition to the raw text message, Tweets carry other metadata that enables the extraction of additional knowledge. Apart from the timestamp that allows us to follow trends and detect emerging topics, the retweeting and reply features for Tweets, and the follower information of a user enable us to learn how information is distributed over the whole social network. Since the members of social networks and their interaction represent nodes and links within a graph, well-know graph analyses can be utilized for knowledge extraction [4].

Apart from knowledge mining related tasks, which could play an important role in e-Participation related projects, we must not forget the functionality of the service itself and how it is used around the world. The recent developments in Tunisia, Egypt and Libya show that social networks were extensively used by the population to communicate, spread news, and organize groups and protests. Although the regimes in these countries tried to block and manipulate the information spread via social networks[4], the processes can still be seen as a major self organizing e-Participation initiative. An earlier example is the utilization of Twitter during the 2009 election in Iran [2].

The basic idea of sentiment analysis is to extract data from Twitter and determine the attitudes towards various subjects and their evolution over time. Due to the wide range of data on Twitter these subjects include things such as products or places, public figures such as politicians or actors, or entities such as companies or discussions about recent events. A good example for the latter one is the discussion about nuclear energy in Germany after the recents events in Fukoshima. Specific examples from current literature are the general discussion about Twitter and sentiment analysis by Go et. al [5], the sentiment analysis of popular terms by Bifet et. al [1], and the prediction of election outcomes in the paper by Tumasjan et. al [14].

Another application-field is the analysis of health-related information. While Quincey et al. [10] discuss the possible application of Twitter for early warning and the detection of pandemics, Rittermann et. al focus on a specific one – the Swine flu pandemic [11]. Another paper within the health sector by Scanfeld et al. analyzes the over-use of antibiotics by extracting information from Twitter [12]. Obviously, another research field is related to the detection of breaking news events or following trends on Twitter [8], [9]. Twitter data has also been used in the financial sector where Wolfram et al. discuss the possibility to use Tweets for modeling the stock market [15]. There has also been an application where the information about published and spread Tweets is used for earthquake detection [3].

This broad range of applications highlights that Twitter is a vital source of information for all kind of data and should definitely be considered in e-Participation related projects.

[4] Attacks on regime critics on Facebook by the Tunisian Goverment: http://www.wired.com/threatlevel/2011/01/tunisia/, Blocking the Internet in Egypt: http://www.nytimes.com/2011/02/21/business/media/21link.html

3 Semantic Patterns (SemPs)

The *Semantic Patterns (SemPs)* technique was developed during the last three years and initially applied to data extracted from the Austrian e-Participation project Mitmachen [13]. In order to identify shortcomings and to improve and extend the method, it was then applied to other domains. These domains include the analysis of malicious code, the correlation of events within Intrusion Detection Systems (IDS), the semantic analysis of RDF data, the investigation of privacy issues within WiFi networks and most recently an automated analysis of metadata extracted from 130.000 applications within the Android market. The application in such heterogenous domains helped us to gain a much better understanding which allowed us to improve the initial technique and integrate it into a Java framework that can be used for the analysis of arbitrary data. Since the in-depth description of the complete technique would go beyond this work, we refer the reader to the previously mentioned publications (especially [7]) for further details.

The main idea behind this technique is to transform a raw data vector containing arbitrary symbolic and real valued features into a pattern, which forms the basis for a wide range of subsequent analyses. This transformation process is depicted in Figure 1 and shows several processing steps that

1. extract terms (nouns, adjectives and verbs), hashtags and timestamps from Tweets and store them as nodes within a semantic network,
2. represent relations between terms, hashtags and timestamps, and the strength of these relations (e.g. defined by the number of co-occurences within a Tweet) as weighted links within this network,
3. apply spreading activation techniques to Tweets, which stimulates the network and spreads the activation of selected nodes according to their links to other regions of the network,
4. and finally extract the activation values for each Tweet from the network and store them within a vector that we call the *Semantic Pattern*.

The generated patterns represent the activation values of different regions within the network that are activated due to different input stimuli (e.g. the hashtag *"#Egypt"* and the term *"protest"*). The distance between two patterns and therefore their similarity can be calculated by the cosine-similarity distance measure. This distance is the basis for a wide range of standard machine learning algorithms.

The key advantages of *SemPs* are the employment of a single, easy-to-interpret model that eliminates the need for complex setups in different domains, and the ability to easily add analysis procedures. Another key advantage is that semantic relations between feature values and not the raw values themselves are stored in the patterns. This removes the need for normalization techniques and enables a straight forward combination of symbolic and real valued features[5].

[5] The data analyzed in this paper only contains symbolic values, but the mixture of symbolic and real values is very typical for other application domains (e.g. the semantic relation between the unemployment rate and an export commodity).

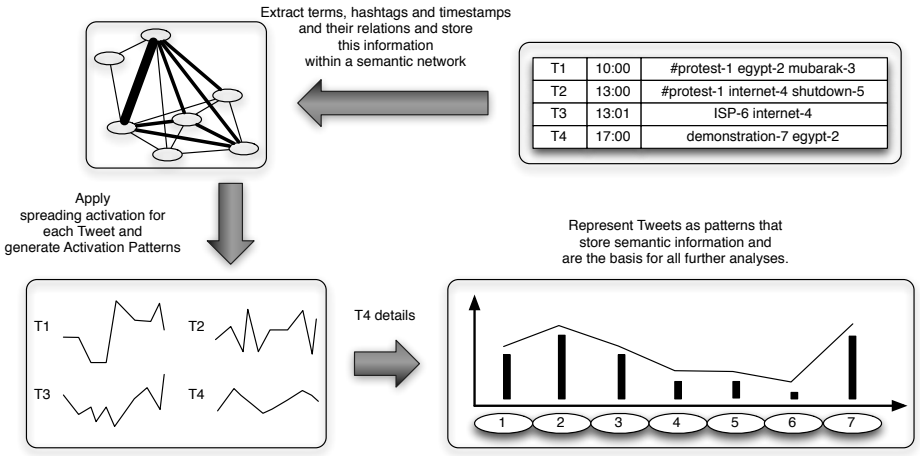


Fig. 1: *SemP* transformation for the Tweets labeled with T1-T4.

4 Accessing Twitter

Although Twitter enables users to communicate via private messages, most of the Tweets are posted to public profiles accessible via a web browser. In addition Twitter offers simple and advanced search interfaces for extracting desired information. The results are chronologically sorted and limited to 1500 Tweets. Especially, when Tweets about hot topics are retrieved, this limit is reached within a few minutes. In addition the real-time results of search queries can also be displayed in widgets that can be added to arbitrary websites. This feature is often used by news sites that display further information from Twitter when covering breaking news events. Furthermore, Twitter offers a streaming API that allows access to the continuous stream of Tweets. This API is the only way to retrieve data which spans a larger timeframe, but comes with the disadvantage that information must already be captured during the unfolding of the monitored events. Therefore, there are several third party services that offer various data to paying customers.

There is also a free alternative that allows the retrieval of Tweets six month back – *Google realtime*[6]. It provides real time search results for social network related data. Although it covers various sources such as Facebook or Twitter, an inspection of the results yields that Twitter is the main data source. The service does not offer any APIs which limits its possible applications. However, since it provides a convenient way to access older Tweets, it still is an interesting alternative. An additional advantage comes with the pre-processing Google applies to the retrieved results. Although there are no specific details on these methods, an empirical analysis suggests that only unique and relevant Tweets

[6] http://www.google.com/realtime

are extracted. This could also be a disadvantage for certain analyses but fitted perfectly for the demonstration presented in this paper.

5 The revolution in Egypt

In this section we demonstrate how the *SemPs* concept can be applied to data extracted from Twitter. Before going into details, we bring the employed knowledge extraction methods into relation with possible e-Participation related projects based on Twitter data. These projects can be assigned to two main categories: The first one is related to projects that ask users to express their opinion on a given topic on Twitter. The topics could simply be separated from other Twitter data by introducing special hashtags. The second category includes projects that use existing Twitter data in order to extract information about arbitrary topics. Some specific examples for such topics are the attitude towards nuclear energy within the last 6 months, the sentiments about infrastructure projects within given regions or the attitude towards political decisions. For both categories we need to extract that Tweets related to the specific topic and use them as basis for subsequent knowledge extraction methods that allow us to draw conclusions on the expressed opinions. Typically, we cannot assume that a-priori knowledge about the analyzed data is available, therefore the applied knowledge extraction methods must enable us to get a good overview of the data and learn key facts before more in-depth analyses can be applied. The *SemPs* concept helps us to achieve this and to avoid the typical problems of setting up domain-specific knowledge extraction methods by using a generic model for a wide range of analyses techniques.

In order to demonstrate the framework and identify key requirements the remainder of this section covers the analysis of Twitter data related to the Egyptian revolution. The data-set was extracted from Google realtime[7] and covers Tweets from January 24^{th} to February 12^{th} 2011. The Tweets have been pre-processed by applying various NLP techniques such as stop-word removal, phrase chunking, part-of-speech (POS) tagging and inflection. Subsequently, the Tweets were parsed and the following three features and their corresponding feature values were extracted: the timestamp of the Tweet, the tokens within the Tweet (nouns, verbs, adjectives) and finally the hashtags. The extracted features and their feature values for each Tweet are the basis for generating the *SemPs* according to the process described in Section 3.

The analyzed data-set was chosen for two main reasons: First, the Egyptian revolution and similar events were also called the social network revolutions since information exchange was carried out over such networks. Therefore these revolutions could be assigned to a special category of e-Participation projects. This was also recognized by the Egyptian government which shutdown the Internet access in response. Secondly, the Egyptian revolution was covered extensively in the news which gives us detailed background knowledge that allows us to verify

[7] One Google realtime search query with the term *"Egypt"*, was executed. The results were parsed via a Java tool and used as input for the *SemPs framework*.

the results of the framework. However – as previously assumed – in the general case such a-priori knowledge is not available. Therefore, a knowledge extraction framework must fulfill these three key requirements:

Analyses layers: The employed algorithms should be able to extract knowledge that allows for an overview of the analyzed data and starting from there going into fine details in subsequent analyses. Further requirements are that the algorithms yield significant results by eliminating noise, and allow for an easy interpretation of these results. If these requirements can be fulfilled with a single model, then the further extension and addition of complex analyses is easier than the application of different algorithms for different tasks. In order to gain a *superficial view* on the analyzed data-set, the most important Tweets must be extracted automatically. The exact definition of important depends on factors like the existing a-priori knowledge of the data-set, the processed data or the desired knowledge. In typical scenarios this comes down to a certain compression or categorization of data. One key technological component here is the application of clustering algorithms. When an overview about the underlying data-set was gained, it is necessary that the analysis framework allows the user to use the superficial analyses as a starting point and go into specific details from there.

Representation: Once data has been extracted, it must be represented to the user. This data representation is a key component, since a bad choice in this area leads to a confusion of the user and cancels the benefits of even the best data extraction algorithms. The appropriate representation depends on the analyzed data, which in the Twitter domain could be a combination or a subset of data such as text, timestamps, or geo-locations. In addition the analyzed data could be of a static or dynamic nature (e.g. an event at a certain timestamp vs. a time frame spanning several months). The representation methods range from simple results lists, over visualizations of time series to maps that either show static or dynamic content.

User interface: The conducted analyses and the representations of the extracted data need to be accessed via a convenient user interface, which is the third key component. This interface must allow to make a seamless transition from layer to layer without the need to execute complex operations. The *Analyses layers* component is already covered with the *Semantic Pattern* concept. However, we are still in the progress of integrating meaningful visualizations, especially for dynamic data, and improving the user interface.

5.1 Getting an overview

For the analyzed Tweets we assume that a-priori knowledge is not available. Therefore, it is crucial that the analysis framework enables the user to gain a quick overview of the data. A common method here is to apply unsupervised learning, or more specific, clustering algorithms that automatically detect categories within the data. Due to the transformation of raw feature values into *SemPs* we are able to directly apply such algorithms. Semantic clustering can be applied to patterns of complete Tweets or to patterns of single feature values (tokens, hashtags, time-stamps) within the Tweets. While the first case is used

to group Tweets covering the same topic into a cluster, the second case can be used to learn more about semantically related feature values (e.g. timestamps for similar events, terms that are used within the same semantic context). In our example this results in the extraction of clusters that cover various topics within the Egyptian revolution such as the protests on Tahrir square, the blocking of the Internet and mobile phone services, the arrest of journalists, or the reported violence during the protests. These clusters help us to gain an overview of the whole data-set and are the basis for further more specific analyses.

The inclusion of the timestamp also enables us to apply clustering algorithms to time series that are generated for Tweets or terms due to semantic changes over a given time frame. When sorting these clusters according to their strongest activity within the time series, we can automatically extract relevant events and arrange them in a timeline. For the analyzed data this includes the following chronologically sorted topics: the accusation of milititants for the bombing of a chruch in December 2010 (2011/01/23), the starting protests (2011/01/25), the following arrests and clashes with the police (2011/01/27), the shutdown of the Internet (2011/01/28), the arresting of journalists, the evacuation of U.S. citizens (2011/02/02), the involvement of the Egyptian army, the final resignation of Hosni Mubarak (2011/02/11), the appointment of an interim military council (2011/02/11) and the international reactions to the revolution.

5.2 Semantic relations of terms, timestamps and hashtags

One key aspect of any analysis is the consideration of *semantic relations* stored within the raw-dataset. This is highlighted via a Tweet that was extracted via the previously mentioned timeline analysis: *"After access is shut down, some ponder if Internet access is a basic human right. 2011/01/29"*. By searching for semantically related Tweets we are able to find more about the incident and other related events. Examples for retrieved Tweets are *"Apparently switching off Twitter is becoming the standard procedure of every country facing social unrest. 2011/01/25"* and *"RT @sharifkouddous I will eventually lose all communication here. But I will be out in the streets tomorrow. 2011/02/01"*. Although these Tweets do not share common terms they are semantically related – meaning they describe similar topics. These relations are domain-specific and typically cannot be transferred to another domain. However, it is still possible to include other domain-invariant information from other knowledge sources (e.g. details about Egypt extracted from DBpedia[8]) that could be used to augment domain-specific semantic relations.

While the semantic relations between terms, Tweets and topics are the most obvious, the concept can be extended to arbitrary data. For this analysis, we also take hashtags and timestamps of Tweets into consideration and link them to the terms within the Tweets. Another Tweet extracted from the timeline analysis reports the following: *"Live footage shows Egypt's army vehicles deploy among protesters at scene of violence in Tahrir square - Al Arabiya TV. 2011/02/03"*.

[8] http://dbpedia.org

By searching for the associated timestamp *"2011-2-3-0"* we can also retrieve other timestamps that are semantically related due to similar events. An example is highlighted by an event that happened on *"2011-2-5-12"*. Here the following Tweet can be retrieved *"Army removing burnt Police vehicle from Tahrir Sq - dark symbol for protestors. 2011/02/05"*. Since for both events similar terms have been used, the corresponding timestamp features are therefore semantically related. The incorporation of this semantic knowledge is the core idea behind the *SemPs* concept and plays a key role for all analyses.

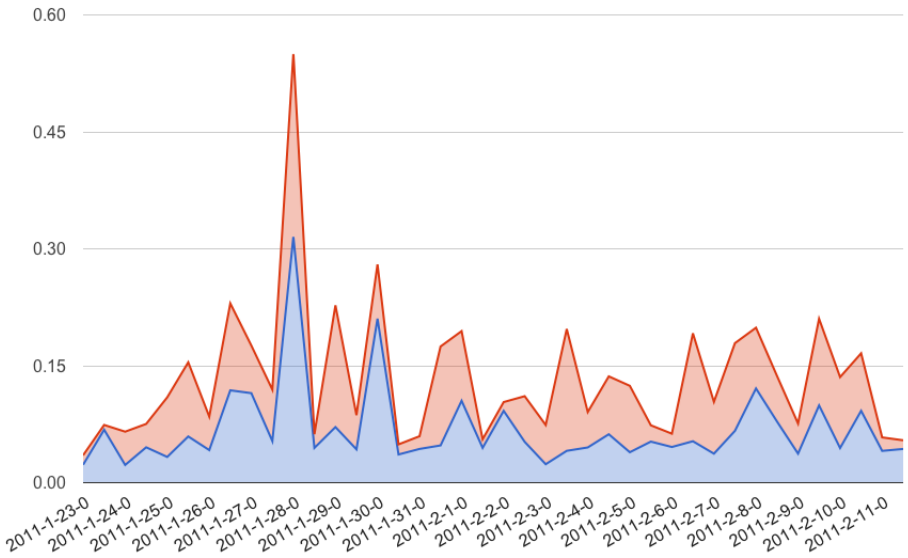
5.3 Going into details

The generated semantic network allows us to extract information about the tokens, hashtags and timestamps stored within the analyzed Tweets. The links and their weight represent the strength of the relations between these features. By using one or more feature values as input we can easily find semantically related information. This information also enables search queries that go beyond simple term matching:

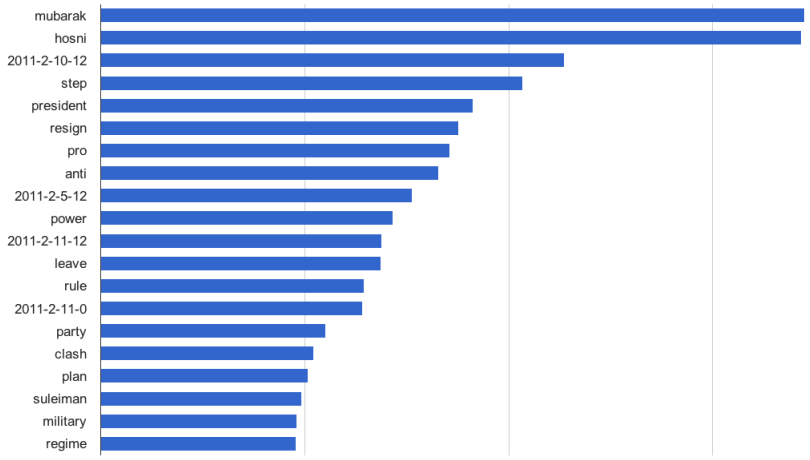
The first example retrieves Tweets that are semantically related to the term *"protest"*. Obviously, Tweets like *"Egypt cracks down on mass protests as Mubarak dissolves government. 2011/1/29"* or *"Egypt anti-government protests start for 3^{rd} day. 2011/1/27"* are the best matching results since they contain the term themselves. Such Tweets could also be retrieved with simple term-matching techniques. However, Tweets like *"Egypt unrest enters third day, El Baradei to return. 2011/01/27"* and *"Journalists now have to register with #Egypt's Ministry of Information if they wish to enter Tahrir, not good. 2011/02/07"* can also be retrieved. They do not contain *"protest"* but other semantically related terms – *"Tahrir", "unrest", and "El Baradei"*.

The second example executes a search query by using the timestamp *"2011-1-25-12"*. At this time the mass protests in Egypt have started. The best matching Tweets are those that were written at that time (e.g. *"Huge protest in Egypt right now as thousands in streets trying to topple Gov't like Tunisia. 2011/01/25"*), but there are also tweets that describe a similar event one day after the first mass protests (timestamp *"2011-03-26-0"*): *"Egypt's Mubarak faces unprecedented protests. Thousands march in the Egyptian capital demanding the end of Hosni. 2011/03/26"*.

The inclusion of the timestamps for each Tweet enables us to generate "semantic time"-patterns that represent the semantic relevance of each feature and Tweet over the complete timeframe. For this data-set twelve-hour intervals were used, which means that a time pattern has roughly 40 entries. By comparing these patterns, one can find Tweets, terms or hashtags that have a similar development over time. As an example we search for Tweets that are related to the event *"Egypt Internet users report major network disruptions. 2011/01/28"*. The retrieved results have a similar activity over time, but do not need to be otherwise semantically related: *"Wikileaks announces it will soon release numerous cables on Egypt. 2011/01/28"* (Figure 2(a)), or *"Egypt protesters, police brace for day of rage. 2011/01/28"*.



(a) Stacked graph for the semantic evolution over time of the Tweets *"Wikileaks announces it will soon release numerous cables on Egypt. 2011/01/28"* (upper graph) and *"Egypt Internet users report major network disruptions. 2011/01/28"* (lower graph). The peak at *"2011-1-28-0"* represents the initial shutdown of the Internet connections and and the peak at *"2011-2-1-0"* represents the shutdown of the last remaining ISP. The y-axis represents the semantic relevance of the Tweets at a given time stamp.



(b) Terms and timestamps that are strongly related with *"Mubarak"* during the revolution. The size of the bars represent the activation values within the semantic network.

Fig. 2: Examples for semantic analyses

The same procedure can be applied to single feature values, which is highlighted by the example *"protest"*: Other terms that have a similar time-pattern are *"police", "tunisia", "government", "people", or "video"*. Although some of these terms are also semantically related, for this example only the time information was utilized.

6 Outlook - Twitter and e-Participation

This paper discusses various application domains for data extracted from Twitter, demonstrates our own knowledge-extraction framework for analyzing Twitter data and based on the learned lessons identifies several key issues that need to be taken into consideration. Based on the findings we strongly argue that Twitter should be used in e-Particpiation related projects and highlight this by drawing the following conclusions: First, due to its huge user base and the continuous coverage of arbitrary topics, we see Twitter as *The Online Presence of our Society* that contains knowledge about arbitrary topics. Second, the automated analysis of Tweets and their carried metadata is vital for the successful extraction of knowledge. Due to the lessons learned from our own analysis framework based on *SemPs* we identify several key requirements for such a knowledge-extraction framework: The semantic knowledge extracted about an arbitrary topic must be presented in several layers that allow the user to make a seamless transition from a superficial overview to fine-grained analyses that extract semantic information. The meaningful representation of the extracted data has a huge impact on the capability of a user to understand important relations and draw further conclusions. Finally, the user interface must allow the user to make smooth transitions between the various analysis layers and address general requirements for intuitive user interfaces. The final conclusion is that Twitter offers the infrastructure for the discussion of topics for free and has a huge user base. Therefore, the future e-Participation projects should consider the possibility to discuss topics directly on the platform. Although there are several disadvantages compared to specific e-Participation related platforms, we argue that the advantages of the huge user base and the ease-of-use outweigh these shortcomings.

Related to our own framework we conclude, that the *SemPs* model represents a well-founded basis that can easily be applied to a wide range of applications and due to its structure can further be extended according to future needs (e.g the inclusion of geo-location based data). Currently, the main target of future improvements are not the analyses layers themselves but the employed data representation layers and the user interface.

References

1. Bifet, A., Frank, E.: Sentiment Knowledge Discovery in Twitter Streaming Data. cswaikatoacnz pp. 1–15 (2010), `http://www.cs.waikato.ac.nz/~eibe/pubs/Twitter-crc.pdf`

2. Burns, A., Eltham, B.: Twitter Free Iran: an Evaluation of Twitter's Role in Public Diplomacy and Information Operations in Iran's 2009 Election Crisis. In: Papandrea, F., Armstrong, M. (eds.) Proceedings of Communications Policy Research Forum. pp. 298–310. Network Insight Institute, University of Technology, Sydney (2009), http://eprints.vu.edu.au/15230/
3. Earle, P.: Earthquake Twitter. Nature Geoscience 3(4), 221–222 (2010), http://dx.doi.org/10.1038/ngeo832
4. Ediger, D., Jiang, K., Riedy, J., Bader, D.A., Corley, C.: Massive Social Network Analysis: Mining Twitter for Social Good. 2010 39th International Conference on Parallel Processing pp. 583–593 (2010), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5599247
5. Go, A., Huang, L., Bhayani, R.: Twitter Sentiment Analysis. Entropy p. 17 (2009), http://nlp.stanford.edu/courses/cs224n/2009/fp/3.pdf
6. Halpin, H., Robu, V., Shepherd, H.: The complex dynamics of collaborative tagging. Proceedings of the 16th international conference on World Wide Web WWW 07 07(1), 211 (2007), http://portal.acm.org/citation.cfm?doid=1242572.1242602
7. Lackner, G., Teufl, P., Weinberger, R.: User Tracking based on Behavioral Fingerprints. In: Proceedings of the The Ninth International Conference on Cryptology And Network Security CANS 2010. p. 0 (2010)
8. Okazaki, M., Matsuo, Y.: Semantic Twitter: Analyzing Tweets for Real-Time Event Notification. Recent Trends and Developments in Social Software 6045, 63–74 (2010), http://www.springerlink.com/index/R40045K20743206J.pdf
9. Phuvipadawat, S., Murata, T.: Breaking News Detection and Tracking in Twitter. 2010 IEEEWICACM International Conference on Web Intelligence and Intelligent Agent Technology pp. 120–123 (2010), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5616930
10. Quincey, E.D., De Quincey, E., Jawaheer, G.: The Potential of Twitter for Early Warning and Outbreak Detection. City (August), 2009–2009 (2009), http://registration.akm.ch/2010eccmid_einsicht.php?XNABSTRACT_ID=102356\&XNSPRACHE_ID=2\&XNKONGRESS_ID=114\&XNMASKEN_ID=900
11. Ritterman, J., Osborne, M., Klein, E.: Using Prediction Markets and Twitter to Predict a Swine Flu Pandemic. Forecast (2004), 1–9 (2009), http://www.iccs.inf.ed.ac.uk/~miles/papers/swine09.pdf
12. Scanfeld, D., Scanfeld, V., Larson, E.L.: Dissemination of health information through social networks: Twitter and antibiotics. American Journal of Infection Control 38(3), 182–8 (2010), http://www.ncbi.nlm.nih.gov/pubmed/20347636
13. Teufl, P., Payer, U., Parycek, P., Macintosh, A., Tambouris, E.: Automated Analysis of e-Participation Data by Utilizing Associative Networks, Spreading Activation and Unsupervised Learning. ePart 09 Proceedings of the 1st International Conference on Electronic Participation 5694, 139–150 (2009)
14. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment (2010), http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1441/1852
15. Wolfram, M.S.A.: Modelling the Stock Market using Twitter. iccsinformaticsedacuk (2010), http://www.iccs.informatics.ed.ac.uk/~miles/msc-projects/wolfram.pdf