SONDY: An Open Source Platform for Social Dynamics Mining and Analysis

Adrien Guille, Cécile Favre
ERIC, University Lyon 2
{adrien.guille,cecile.favre}@univ-lyon2.fr

Hakim Hacid
Alcatel-Lucent Bell Labs France
hakim.hacid@alcatel-lucent.com

Djamel Zighed
Institute of Human Science, University Lyon 2
abdelkader.zighed@ish-lyon.cnrs.fr

ABSTRACT

This paper describes *SONDY*, a tool for analysis of trends and dynamics in online social network data. *SONDY* addresses two audiences: (i) end-users who want to explore social activity and (ii) researchers who want to experiment and compare mining techniques on social data. *SONDY* helps end-users like media analysts or journalists understand social network users interests and activity by providing emerging topics and events detection as well as network analysis functionalities. To this end, the application proposes visualizations such as interactive time-lines that summarize information and colored user graphs that reflect the structure of the network. *SONDY* also provides researchers an easy way to compare and evaluate recent techniques to mine social data, implement new algorithms and extend the application without being concerned with how to make it accessible. In the demo, participants will be invited to explore information from several datasets of various sizes and origins (such as a dataset consisting of 7,874,772 messages published by 1,697,759 Twitter users during a period of 7 days) and apply the different functionalities of the platform in real-time.

Categories and Subject Descriptors

H.3.4 [**Information Systems**]: Information Storage and Retrieval—*Systems and Software*

Keywords

Online social networks, topic detection, network analysis

1. INTRODUCTION

Online social networks allow hundreds of millions of Internet users worldwide to produce and consume content. They provide access to a very vast source of information on an unprecedented scale. Still, the raw data produced by users of these networks is a flood of ideas, information, opinions, *etc.* Given the impact of online social networks on society and the strategic interest for industries, the recent focus is on extracting valuable information from this huge amount of "social" data. Events, issues, interests, *etc.* happen and evolve very quickly in social networks and their capture, understanding, visualization, and prediction are becoming critical expectations from both end-users and researchers. This is motivated by the fact that understanding the dynamics of these networks may help in better following events, solving issues (*e.g.* natural hazards), anticipating needs (*e.g.* new products), *etc.*

Dynamics analysis in social networks follows generally similar principles: given a subject of interest such as a set of individuals or topics, the idea is to observe and analyze their evolution over time according to specific metrics. Most of the time, actions are taken afterwards on the subjects according to underlying users expectations. Fully understanding and capturing the dynamics in social networks is a very hard task due to the complexity of the social structures in terms of size, relations, content heterogeneity, *etc.* In fact, a large portion of social content is highly unstructured, most of the time messy, and disparate. Although several contributions exist towards dynamics analysis in social data, most of them don't provide implementations of their techniques, and the few existing implementations are written in different languages and require different formatting and preparation of the input data, making it nearly impossible to compare and experiment with the various approaches. Besides the difficulties of developing new techniques for topic detection, these tasks necessitate generally a heavy preprocessing step which is performed manually.

We propose in this work *SONDY*[1] (*i.e. SOcial Network DYnamics*), a tool that tackles the following two issues: (i) how to assist researchers and end-users in preprocessing data from online social networks, detecting topics and their trends, analyzing the corresponding networks (*i.e.* active authors for the considered topic(s)), and (ii) how to make it effortless to integrate, compare, and eventually combine different approaches to mine such data. *SONDY* is an open source platform integrating optimized implementations of some topic detection and graph mining algorithms in the same platform. The application relies on four services to address the mentioned issues:

1. *Data manipulation service*: for importing and preparing the data in order to optimize their exploitation and processing. This component includes stop-words removal, content stemming, message stream discretization, and message stream resizing.

[1] *http://mediamining.univ-lyon2.fr/sondy/*

SIGMOD'13, June 22–27, 2013, New York, New York, USA.

2. *Topic detection and exploration service*: for identifying and temporally locating trending topics and events. It encapsulates a set of configurable algorithms for trends detection combined with results visualization under several customizable forms.

3. *Network analysis and visualization service*: for observing the social network structure and finding, *e.g.* influential nodes or communities. Visualizations are interactive, making it possible for users to actively interact with the system.

4. *Extension manager*: for importing new algorithms to be used by the the topic detection or network analysis services.

To illustrate the capabilities of *SONDY*, the demonstration will consist in inviting participants to, on-live, (i) operate some preprocessing on large Twitter datasets, select time periods, (ii) apply detection algorithms and modify their parameters in order to explore, compare and visualize trending topics and (iii) analyze the structure of the network of authors of the messages w.r.t a selected topic and time-period. This will be done in an interactive manner through the tool's user interface. Finally, users will be shown how easy it is to integrate new algorithms into the platform. To do so, an already implemented algorithm will be used as an example.

2. PLATFORM DESIGN

In this section, we describe the different components of *SONDY* to highlight their aim and the way they cooperate.

2.1 Architecture

SONDY offers four main services to ensure an effortless and complete analysis of social dynamics and comparison of data mining techniques as described in the previous section. It is written in Java (about 10K lines of code) because of its ease of use and high reliability. In addition to the offered services, *SONDY* provides an easy way for researchers to implement new algorithms and integrate them into the application using a programming interface without spending time in, *e.g.* data manipulation or visualization considerations since they are managed natively in *SONDY*. New algorithms are added using the *extension manager service*. The different services are described in the next sections and the way they communicate together is represented in Figure 1.

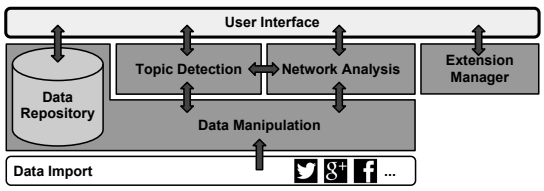

Figure 1: Overview of the architecture of *SONDY*.

2.2 Data Manipulation

This service manages a collection of datasets (Figure 2.a) and offers not only the ability to import new social content in the collection (Figure 2.b) but also a set of filters to prepare it for further processing (Figure 2.c). The basic input for import is composed of a set of messages and, if it is available, the network of relationships of the related authors. Messages are composed of a free text, an author and a timestamp. When a new dataset is imported in the collection, the application stores it in an ad-hoc indexed database[2]. This is a crucial step to get high performance especially when dealing with large datasets. From there, the dataset can be preprocessed with the following filters:

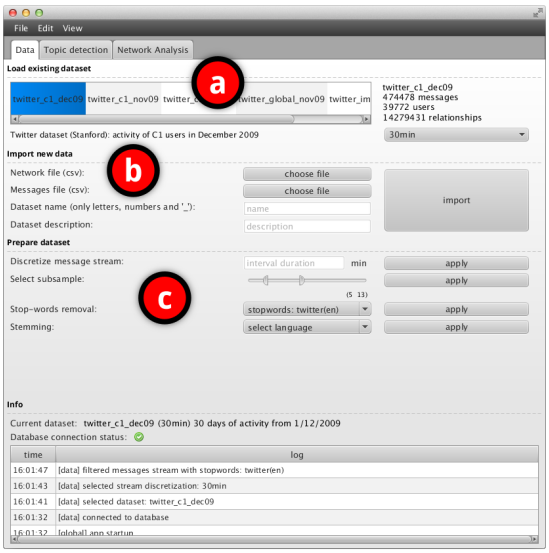

Figure 2: Data import and preprocessing service: on this illustration, we see how to resize the message stream and remove specific twitter stop-words.

- *Message stream discretization filter:* slices the message stream according to a chosen window of time, so that detection algorithms based on term frequency can be applied. During this step, messages content is indexed using the *Lucene* API[3] for further retrieval.
- *Message stream resizing filter:* allows the user to select a temporal subsample of the message stream on which the analysis should focus.
- *Stop-words filter:* removes stop-words from messages, based either on one of the lists provided with the application or a customized list provided by the user.
- *Stemming filter:* reduces words to their stem to improve efficiency of topic detection algorithms, such as topic model based techniques.

2.3 Topic Detection and Exploration

This service enables (i) applying different topic detection algorithms on the same dataset (Figure 3.a) and (ii) interactively exploring the trends of the detected topics: by exploring the ranked result table (Figure 3.b) in which one can search for specific terms (Figure 3.c), by exploring related messages (Figure 3.d), by plotting and comparing terms usage (Figure 3.e), or by generating time-lines (Figure 3.f) to

[2]We use a *MySQL* server *http://www.mysql.com* for this version of the tool.
[3]*http://lucene.apache.org*

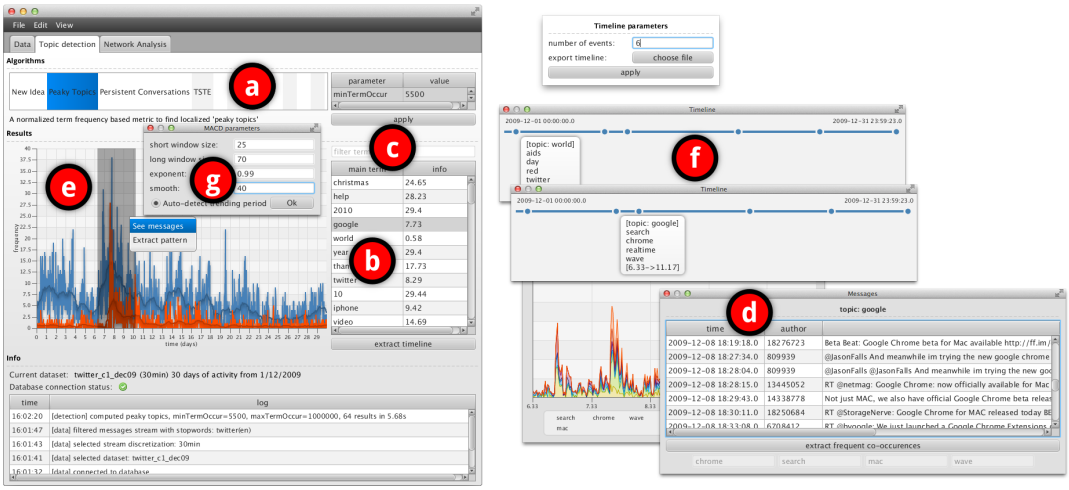

Figure 3: Illustration of the different components offered by the trends detection and exploration service.

summarize results. So far we have implemented the following topic detection and scoring metrics: *Peaky Topics* [8], *Persistent Conversations* [8], and the *Temporal and Social Term Evaluation* (*TSTE*) [4].

The *Peaky Topics* metric detects and ranks highly localized and momentary terms of interest while the *Persistent Conversations* metric ranks less salient terms which sustain for a longer duration using a modified version of the *tf-idf* metric adapted to message streams. *TSTE* is based on the exploitation of term frequency and authors influence. The authority of the active authors is assessed using their relationships and the *PageRank* algorithm [6]. It allows to model the life cycle of each term on the basis of a biological metaphor, which is based on the calculation of values of nutrition and energy that leverage the users authority. It relies on the computation of a critical drop value based on the energy, to select sets of emerging terms and rank them.

We also implemented the *Moving Average Convergence Divergence* (*MACD*) indicator [7]. It helps identify periods of time (highlighted on the plot) during which terms are trending (Figure 3.g). The principle of *MACD* is to turn two trend-following indicators, precisely a short period and a longer period moving average of term frequency, into a momentum oscillator.

Figure 4: Visualizing a network of authors for a selected topic and time period, colored with k-shells.

2.4 Network Analysis and Visualization

To assist in the understanding of how trends emerge, this service proposes to visualize the network of authors about a topic, more precisely the active authors for the selected topic and time period in the detection service, in order to, *e.g.* identify influential nodes or detect communities.

To improve visualizations, *SONDY* currently implements two techniques for influence analysis: (i) *k-shell decomposition* [2] and (ii) *PageRank* [6]. The *k-shell decomposition* is a powerful tool for the identification of influential spreaders in complex networks. It consists in identifying particular subsets of the graph, called k-cores, each one obtained by recursively pruning least connected nodes, *i.e.* all the nodes of degree smaller than k, until the degree of all remaining nodes is larger than or equal to k. Larger values of k (*i.e.* "coreness") correspond to nodes with larger degree and more central position in the network structure. We implemented this method with an $O(n)$ algorithm, where n is the number of arcs. *PageRank* [6] is a well known technique for measuring the importance of nodes in any given graph. The *PageRank* value of a node v is proportional to the probability of visiting v in a random walk of the social network, where the set of states of the random walk is the set of nodes. The probability of being at each node (*i.e.* state) is computed using an iterative update method that is guaranteed to converge.

The results of these algorithms (Figure 4.a) are shown on colored graphs (Figure 4.b), using the *Graphstream*[4] API, and plots representing nodes color distribution over the network (Figure 4.c). The graph visualization is interactive

[4] *http://graphstream-project.org/*

and allows the identification of nodes and exploration of the messages they published.

2.5 Extension Manager

SONDY is an open source application that allows researchers to experiment and compare different methods and thus offers a programming interface for simple integration of new algorithms. Algorithms need to conform to a particular method definition, declare their parameters so they can be set through the UI, and use the data manipulation functions provided by the application. Once this is done, the new algorithm can be simply given as input for the extension manager service, under the form of a JAR file, which will make it accessible through the user interface.

3. DEMO DESCRIPTION

In this demo, participants will be invited to explore information from several datasets of various sizes and origins (such as a dataset consisting of 7,874,772 messages published by 1,697,759 Twitter users during a period of 7 days). Our objective here is to illustrate that it is almost effortless and very efficient to manage datasets and filter them using *SONDY*. Participants will be able to select any existing dataset, resize and filter it. The audience will then be able to experiment with the different algorithms and compare the detected topics, depending on the previously selected discretization and filters. They will be able to explore messages concerning the topic and time period of their choice, compare computation times of the various techniques, the number of detected topics, *etc.* Participants will also be able to visualize the network structure highlighted using *k-shell decomposition* or *PageRank* for the topic and time period they selected, in order to identify, *e.g.* influential spreaders, and how the distribution of influence varies according to topics and active authors. To ensure a comfortable viewing, the visualizations can be made full screen and stay interactive. Finally, users will be shown how easy it is to integrate new algorithms into the platform. To do so, an already implemented algorithm will be used as an example.

4. RELATED WORK

To the best of our knowledge, there isn't yet a tool for analyzing the whole social dynamics, that is to say, analyzing both social activity via the exploitation of messages and networks in relation with the activity. *SONDY* intends to bridge this gap.

Several tools for network analysis and graph mining have been developed. *Cuttlefish*[5] offers dynamic network visualization and simulations using different layouts. *Gephi*[6] is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs. *SNAP*[7] is a C++ library developed at Stanford that can scale to massive networks and calculates structural properties.

Some tools for topic detection and visualization have been developed in the recent years. They mainly suffer from two issues: (i) they are designed to work with a specific data source and (ii) they don't permit to use other algorithms than the one they implement. *TwitInfo* [5] tracks keyword mentions on Twitter and turns them into time-line visualizations which summarize various topics using an ad-hoc algorithm. *Eddi* [3] offers visualizations such as time-lines and tag clouds of topics extracted from tweets using a simple topic detection algorithm that uses a search engine as an external knowledge base.

[5] *http://cuttlefish.sourceforge.net/*
[6] *http://gephi.org/*
[7] *http://snap.stanford.edu/*

5. CONCLUSION AND FUTURE WORK

We presented in this paper *SONDY*, an open source tool for analyzing social dynamics from social networks data and comparing mining techniques. The tool is designed for end-users and researchers. *SONDY* helps end-users like media analysts or journalists understand social networks users interests and activity by providing emerging topics and events detection as well as influence analysis functionalities. To this end, the application proposes visualizations such as interactive time-lines that summarize information and colored users graph that reflect the structure of the network. This is the first tool bridging the gap between content and structure analysis in social networks. *SONDY* also provides researchers an easy way to compare and evaluate recent techniques to mine social data, implement new algorithms and extend the application without being concerned with how to make it accessible. As future work, we intend to enrich the available range of techniques for topic detection, trends analysis, as well as network analysis to offer a larger panoply of algorithms for comparison. Algorithms include for example *Online LDA* [1] for topic detection and community detection techniques for network analysis. To increase performances, we plan to incorporate a *Hadoop*[8] support for larger datasets and parallel processing.

6. REFERENCES

[1] L. AlSumait, D. Barbará, and C. Domeniconi. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *ICDM '08*, pages 3–12, 2008.

[2] V. Batagelj and M. Zaversnik. Fast algorithms for determining (generalized) core groups in social networks. *Advances in Data Analysis and Classification*, 5(2):129–145, 2011.

[3] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi. Eddi: interactive topic-based browsing of social status streams. In *UIST '10*, pages 303–312, 2010.

[4] M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *MDMKDD '10*, pages 4–13, 2010.

[5] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *CHI '11*, pages 227–236, 2011.

[6] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *WWW '98*, pages 161–172, 1998.

[7] L. Rong and Y. Qing. Trends analysis of news topics on twitter. *International Journal of Machine Learning and Computing*, 2(3):327–332, 2012.

[8] D. A. Shamma, L. Kennedy, and E. F. Churchill. Peaks and persistence: modeling the shape of microblog conversations. In *CSCW '11*, pages 355–358, 2011.

[8] *http://hadoop.apache.org/*