Exploring the Wisdom of the Tweets: Towards Knowledge Acquisition from Social Awareness Streams

Claudia Wagner

JOANNEUM RESEARCH,
Steyrergasse 17, 8010, Graz, Austria
claudia.wagner@joanneum.at

Abstract. Although one might argue that little wisdom can be conveyed in messages of 140 characters or less, this PhD research sets out to explore if and what kind of knowledge can be acquired from different *aggregations of social awareness streams*. The expected contribution of this research is a network-theoretic model for defining, comparing and analyzing different kinds of social awareness streams and an experimental prototype to extract semantic models from them.

Key words: Knowledge Acquisition, Social Web, Semantic Analysis

1 Introduction

In the last decade, the emergence of social media applications such as Wikipedia, Del.icio.us and Flickr has inspired a community of researchers to tap into user-generated data as an interesting alternative to knowledge acquisition. Instead of formally specifying meaning *ex-ante* through for example agreed-upon ontologies, the idea was to capture meaning from user-generated data *ex-post*.

With the emergence of social awareness streams, popularized by applications such as Twitter or Facebook and formats such as activitystrea.ms, a new form of communication and knowledge sharing has enriched the social media landscape. Personal awareness streams usually allow users to post short, natural-language messages as a personal stream of data that is being made available to other users. We refer to the aggregation of such personal awareness streams as *social awareness streams* (short streams), which contain a set of short messages from different users usually displayed in reverse chronological-order. Although one could argue that little wisdom can be conveyed in messages of 140 characters or less, this PhD research aims to explore (1) if and what kind of knowledge can be acquired from different *aggregations of social awareness streams* and (2) to what extent the semantics of social awareness streams are influenced by stream characteristics and vice versa.

2 Related Work

Semantic analysis of social media applications is an active research area, in part because on the one hand social media provides access to the "collective intelligence" of millions of users while on the other hand it lacks explicit semantics. Exploiting the "collective intelligence" of social media applications is therefore a promising and challenging aim of current research efforts.

The work of Mika [1] and Heymann et al [2] illustrate how graph-based measures, such as centrality and clustering coefficient, can be used to extract broader and narrower terms from tag spaces. Schmitz et al. [3] describe how a statistical subsumption model can be applied to induce hierarchical relations of tags. Clustering approaches (e.g., [4]) identify groups of highly related tags and establish hierarchical relationship between these clusters.

Although social awareness streams and tagging systems have common characteristics (e.g., in both systems users relate resources with free-form tags), they are used for different purpose and reveal significant, structural differences (e.g., in social awareness streams users may relate resources with other resources or other users). Due to its novelty, little research on social awareness streams exists to date. Some recent research (e.g., [5]) investigates user activities on Twitter. Another line of research (e.g., [6]) focuses on analyzing and characterizing content of social awareness stream messages.

3 Proposed Approach and Methodology

To characterize and compare different aggregations of streams we will develop a network-theoretic model and measures for social awareness streams. To explore if and what kind of knowledge can be acquired from streams, a system (KASAS) for characterizing and comparing stream aggregations and extracting emerging semantic models from them will be developed.

3.1 The KASAS system

Figure 1 shows the basic steps of the KASAS system which takes one or several keywords as input and returns as output a model of concepts and relations between them. The resulting model could for example be used to enrich existing ontologies such as DBpedia with semantic models (containing e.g. recent information about events or users relevant for a specific DBpedia concept). The **Stream Aggregation and Characterization** component creates for a given keyword one or several aggregations of streams and characterizes them via various stream measures. These measures can e.g. help to identify the most promising streams in terms of semantic analysis. The streams are then preprocessed by the **Lexical Normalization** component. Concepts (denoted by one or several labels) and their relations are extracted by the **Concept and Association Mining** component. This component will use simple network transformations, latent and explicit concept models to extract concepts from stream aggregations. To mine

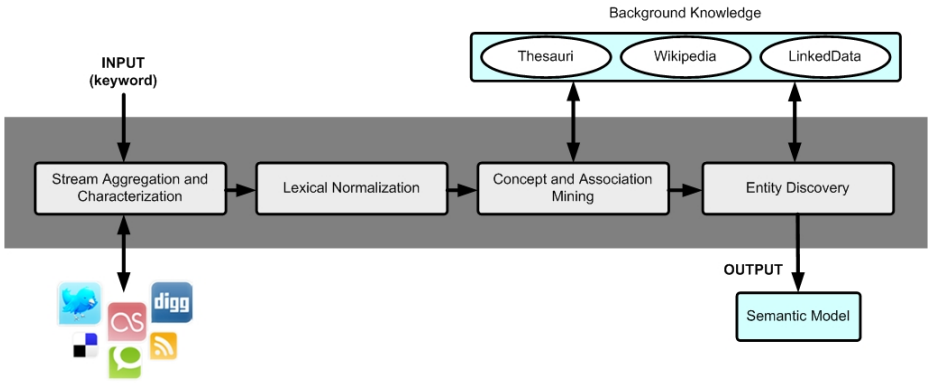

Fig. 1. The KASAS (**K**nowledge **A**cquisition from **S**ocial **A**wareness **S**treams) system

associations between concepts we will use various information-theoretic, statistical, semantic similarity measures. Finally, the **Entity Discovery** component will discover types of concepts by harnessing the Web of Data as background knowledge and by analyzing stream characteristics. Stream characteristics can help to infer the basic type of a stream's topic, which is e.g. `event` in case of the `#eswc2010` stream.

3.2 Evaluation

When it comes to the semantic analysis of social awareness streams, the extent to which different streams approximate the semantic understanding of users participating in these streams is interesting to investigate. We will conduct evaluations that include user assessments of concepts, relations and their ranking. In addition, to evaluate the quality of extracted concepts and their relations, we may use external, hand-crafted taxonomies, such WordNet, the Open Directory Project or DBpedia as semantic grounding, and internal, semantic models emerging through user's usage of special syntax, such as microsyntax[1] or HyperTwitter[2].

To evaluate concept-user relations we can select a defined set of user accounts belonging to researchers and compare their list of ranked concepts with a list of keywords of their papers. In addition, we may ask themselves (self-assessment) and other researchers (peer-assessment) to asses the extracted and ranked list of concepts related with them.

4 Results

Based on the existing tripartite structure of *folksonomies*, we introduce a network-theoretic model of social awareness streams consisting of messages, users and

[1] `http://microsyntax.pbworks.com/`
[2] `http://semantictwitter.appspot.com/`

content of messages. As an adaption of the folksonomy data structure, the model of social awareness streams introduces qualifiers on the tripartite structure that allow to accommodate user generated syntax. We formally define the model as follows:

Definition 1 *A social awareness stream S is a tupel $S = (U_{q1}, M_{q2}, R_{q3}, Y, ft)$, where*

- *U, M and R are finite sets whose elements are called users, messages and resources.*
- *Qualifier q1 represents different ways users are involved in a stream (e.g. users can be author or target of message), q2 represents the different types of messages M (e.g. public broadcast messages or private direct messages), and q3 represents the different types of resources that can be included in messages of streams (e.g. hashtags, links or keywords)*
- *Y is a ternary relation $Y \subseteq U \times M \times R$ between U, M, and R.*
- *ft is a function which assigns to each Y a temporal marker.*

In addition, we created various stream measures (such as the social, conversational, temporal, informational and topical diversity measure) to characterize and compare different stream aggregations. Based on the formal model of social awareness streams and the predefined measures, we analyzed and compared different aggregations of Twitter streams (user list, user directory, hashtag and keyword streams), which were all related to the concept `semantic web` and were all recorded within the same time interval (8 days).

Since measures for similarity and relatedness are not well developed for three-mode networks we considered various ways to obtain qualified two-mode networks (resource-author, resource-message, resource-hashtag and resource-link networks) from these stream aggregations. To surface semantic relations between resources, we produced one-mode networks (see e.g. Figure 2) by multiplying the corresponding two-mode network matrices with their transpose $M * M^T$. Different semantic relations are created because of the different ways associations are established between resources. We compared the resulting networks by assessing their most important concepts and relations. Our empirical results indicate that hashtag streams are in general rather robust against external events (such as New Years Eve), while user list stream aggregations are more perceptible to such "disturbances". Networks generated via resource-hashtag transformations seem to have the power to reduce the non-informational noise in streams and reveal meaningful semantic models.

5 Conclusions and Future Work

While the developed network-theoretic model of social awareness streams is general, the first empirical results of this PhD research are limited to a single concept (`semantic web`). In future, we will expand our analysis to a broader variety of social awareness streams and conduct experiments over greater periods of time.

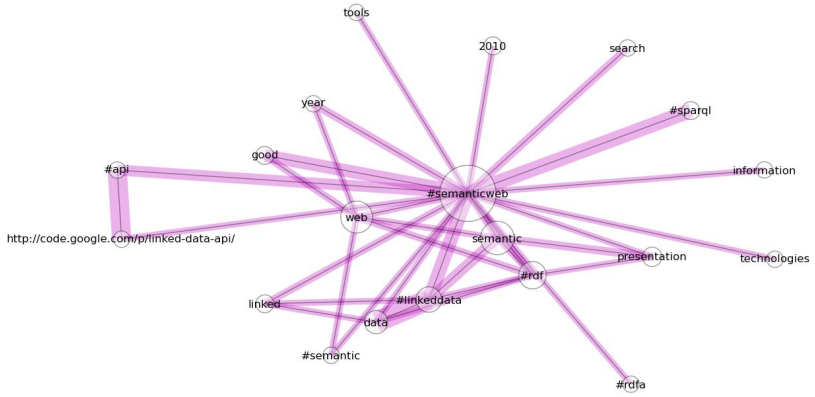

Fig. 2. Resource network computed from the resource-hashtag network of the *#semanticweb* hashtag stream

Since the semantic analysis conduced in our first experiments is based on simple network transformations, we will study whether more sophisticated knowledge acquisition methods produce different results. Finally, we will evaluate the semantic models produced by different stream aggregations and explore to what extend they approximate the semantic understanding of users that are involved in these streams.

Acknowledgments. I would like to thank my supervisor Markus Strohmaier for his guidance, support and fruitful discussions and JOANNEUM RESEARCH for funding this research.

References

1. Mika, P.: Ontologies are us: A unified model of social networks and semantics. Web Semant. **5**(1) (2007) 5–15
2. Heymann, P., Garcia-Molina, H.: Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Computer Science Department (April 2006)
3. Schmitz, P.: Inducing ontology from flickr tags. In: Proceedings of the Workshop on Collaborative Tagging at WWW2006, Edinburgh, Scotland (May 2006)
4. Begelman, G., Keller, P., Smadja, F.: Automated tag clustering: Improving search and exploration in the tag space. In: Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland. (2006)
5. Huberman, B.A., Romero, D.M., Wu, F.: Social networks that matter: Twitter under the microscope. ArXiv e-prints (December 2008)
6. Naaman, M., Boase, J., Lai, C.H.: Is it all about me? user content in social awareness streams. In: Proceedings of the ACM 2010 conference on Computer supported cooperative work. (2010)