The Wisdom in Tweetonomies: Acquiring Latent Conceptual Structures from Social Awareness Streams

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

Markus Strohmaier
Graz University of Technology and Know-Center
Inffeldgasse 21a
8010 Graz, Austria
markus.strohmaier@tugraz.at

ABSTRACT

Although one might argue that little wisdom can be conveyed in messages of 140 characters or less, this paper sets out to explore whether the *aggregation of messages* in social awareness streams, such as Twitter, conveys meaningful information about a given domain. As a research community, we know little about the structural and semantic properties of such streams, and how they can be analyzed, characterized and used. This paper introduces a network-theoretic model of social awareness stream, a so-called "tweetonomy", together with a set of stream-based measures that allow researchers to systematically define and compare different stream aggregations. We apply the model and measures to a dataset acquired from Twitter to study emerging semantics in selected streams. The network-theoretic model and the corresponding measures introduced in this paper are relevant for researchers interested in information retrieval and ontology learning from social awareness streams. Our empirical findings demonstrate that different social awareness stream aggregations exhibit interesting differences, making them amenable for different applications.

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 or taxonomies, 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*, which usually contain a set of short messages from different users. Although one could argue that little wisdom can be conveyed in messages of 140 characters or less, this paper sets out to explore whether the *aggregation of messages* in different social awareness streams conveys meaningful information about a given domain.

Extracting structured knowledge from unstructured data is a well-known problem which has extensively been studied in the context of semantic search, because semantic search attempts to consider the meaning of users‘ queries and of available web resources. To extract the meaning of available web resources, different methods have been proposed which mainly rely on the content of web pages, their link structure and/or collaboratively generated annotations of web pages, so-called folksonomies. Social awareness streams provide a rich source of information, which can for example be used to improve semantic search by revealing possible meanings of a user's search query and by providing social annotations of web resources.

Since social awareness streams differ significantly from other information sources, such as web pages, blogs and wikis (e.g., through their lack of context and data sparseness), chats and newsgroups (e.g., through the way how information is consumed on social awareness streams, namely via social networks) and social tagging systems (e.g., through their structure and purpose), their applicability for knowledge acquisition and semantic search is still unclear. To address these differences and capture information structures emerging from social awareness streams we introduce the concept of a "tweetonomy", a three-mode network of social awareness streams.

This paper sets out to explore characteristics of different *social awareness stream aggregations* and analyzes if and what kind of knowledge can be extracted from social awareness streams through simple network transformations. The overall objectives of this paper are 1) to define a network-theoretic model of social awareness streams that is general enough to capture and integrate emerging usage syntax, 2) to define measures that characterize different properties of social awareness streams and 3) to apply the model together with the measures to study semantics in Twitter streams. Our experimental results show that different types of social awareness streams exhibit interesting differences in terms of the semantics that can be extracted from them. Our findings have implications for researchers interested in ontology learning and information retrieval from social awareness streams or general studies of social awareness streams.

The paper is organized as follows: First we introduce a network-theoretic model of social awareness streams as a tripartite network of users, messages and resources. Then, we propose several measures to quantify and compare differ-

WWW2010, April 26-30, 2010, Raleigh, North Carolina.
.

ent properties of social awareness streams. Subsequently, we characterize four different types of social awareness streams which have been aggregated from Twitter for a given search query *semantic web*, by computing several structural stream measures, such as the social and topical diversity of a stream. We investigate if and what kind of knowledge can be acquired from different aggregations of social awareness streams by transforming them into lightweight, associative resource ontologies. Finally, we relate our work to other research in this area and draw conclusions for future work.

2. SOCIAL AWARENESS STREAMS

Social awareness streams are an important feature of applications such as Twitter or Facebook. When users log into such systems, they usually see a stream of messages posted by those they follow in reverse chronological order. That means information consumption on social awareness streams is driven by explicitly defined social networks. Although messages in social awareness streams can be targeted to specific users, they are broadcasted to everyone who follows a stream and can be public or semi-public (i.e., only visible to users belonging to a user's social network).

Messages usually consist of words, URLs, and other user-generated syntax such as hashtags, slashtags or @replies. Hashtags are keywords prefixed by a hash (#) symbol which enrich short messages with additional (often contextual) information. Hashtags are, amongst others, used to create communication channels around a topic or event and to annotate term(s) with additional semantic metadata (e.g., #need[list of needs][1]). Slashtags[2] are keywords prefixed by a slash symbol (/) to qualify the nature of references in a message. So called @replies are usernames prefixed by an at (@) symbol and are used to mention users or target messages to them.

In contrast to other stream-based systems where data structures are formally defined by system developers (such as the tripartite data structure of folksonomies), social awareness streams are different in the sense that they have yielded an emerging, collectively-generated data structure that goes far beyond what the system designers' have envisioned. Emerging syntax conventions, such as RT (retweets), # (hashtags) or @ (replies), are examples of innovations by users or groups of users that superimpose an informal, emerging data structure on social awareness streams. This has made social awareness streams complex and dynamic structures which can be analyzed in a staggering variety of ways, for example, according to the author(s) of messages, the recipients of messages, the links, keywords or hashtags contained in messages, the time stamps of messages or the message types.

2.1 Tweetonomy: A Tripartite Model of Social Awareness Streams

Based on the existing tripartite structure of *folksonomies* [14] [7] [16] [5], we introduce a tripartite model of social awareness streams, a so-called "tweetonomy", which consists of messages, users and content of messages.

While a taxonomy is a hierarchical structure of concepts developed for classification, a folksonomy refers to the emerging conceptual structure that can be observed when a large group of users collaboratively organizes resources. In a tweetonomy nobody classifies or organizes resources, but users engage in casual chatter and dialogue. Our motivation for introducing *tweetonomies* as a novel and distinct concept is rooted in our interest in knowledge acquisition from this new and different form of discourse, i.e. to explore whether we can acquire latent hierarchical concept structures from social awareness streams such as Twitter of Facebook.

To formally define emerging structures from social awareness streams we present the model of a tweetonomy and introduce qualifiers on the tripartite structure that allow to accommodate user generated syntax. We formally define a tweetonomy as follows:

Definition 1. *A tweetonomy is a tupel* $T := (U_{q1}, M_{q2}, R_{q3}, Y, ft)$*, where*

- *U, M and R are finite sets whose elements are called users, messages and resources.*
- *Qualifier q1 represents the different ways in which users can be related to a message. For example, a user can be the author of a message (U_a), or a user can be mentioned in a message in a variety of ways, such as being mentioned via an @reply ($U_@$), or being mentioned via slashtags[3] such as /via, /cc and /by, which can represented as U_{via},U_{cc} and U_{by}. Future syntax can be accommodated in this model by adding further types of relations between users and messages.*
- *Qualifier q2 represents the different types of messages M supported by a social awareness stream. Messages in social awareness streams can have different qualities depending on the system. For example, the Twitter API distinguishes between public broadcast messages (M_{BC}), conversational direct messages (M_D), and retweeted messages (M_{RT}). Future syntax can be accommodated in this model by adding further message types.*
- *Qualifier q3 represents the different types of resources that can be included in a social awareness stream. Resources can be keywords (R_k), hashtags (R_h), URLs (R_l) or other informational content occurring in messages of a social awareness stream.*
- *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,* $ft : Y \rightarrow \mathbb{N}$.

If we mention U, M or R without any qualifier, we refer to the union of all qualified sets of them. According to the definition, we use U_a to refer to the set of users who authored messages of stream, U_m to refer to the set of users who are mentioned in messages of a stream and R_h, R_k and R_l to refer to the set of resources in messages which are hashtags, keywords and URLs.

To define and characterize social awareness streams as well as individual messages, we can use the tripartite model to represent them as a tuples of users, messages and resources. For example, the following Twitter message: "*RT@tim new blog post: http://mydomain.com #ldc09*" created by a user *alex* can formally be represented by the tweetonomy shown in Figure 1.

[1] http://epic.cs.colorado.edu/helping_haiti_tweak_the_twe.html

[2] http://factoryjoe.com/blog/2009/11/08/

[3] http://factoryjoe.com/blog/2009/11/08/

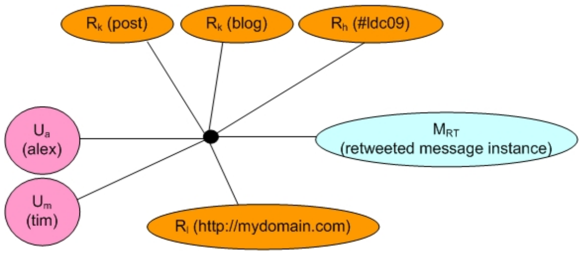


Figure 1: Example of a simple tweetonomy

2.2 Aggregations of Social Awareness Streams

The tripartite structure provides a general model to distinguish different aggregations of social awareness streams. Depending on the task and scope of investigations, researchers usually have to make choices about which aspects of social awareness streams to study. By making these choices, they usually produce different aggregations of the stream of data, that capture different parts and dynamics of streams. The introduced tripartite model allows to make these choices explicit.

In the following, we use the tweetonomy model to 1) define a subset of different aggregations of social awareness streams and 2) to demonstrate the nature and characteristics of different aggregations. Based on the model, we can distinguish between three basic aggregations of social awareness: resource streams $S(R')$, messages streams $S(M')$ and user streams $S(U')$. They are defined in the following way:

A resource stream $S(R')$ is a tupel $S(R') = (U, M, R, Y', ft)$, where $Y' = \{(u, m, r) \,|\, r \in R' \vee \exists r' \in R', \tilde{m} \in M, u \in U : (u, \tilde{m}, r') \in Y\}$ and $R' \subseteq R$ and $Y' \subseteq Y$. In words, a resource stream consists of all messages containing one or several specific resources $r' \in R'$ (e.g. a specific hahstag, URL or keyword) and all resources and users related with these messages.

A user stream $S(U')$ is a tupel $S(U') = (U, M, R, Y', ft)$, where $Y' = \{(u, m, r) \,|\, u \in U' \vee u' \in U', \tilde{m} \in M, r \in R : (u', \tilde{m}, r) \in Y\}$ and $U' \subseteq U$ and $Y' \subseteq Y$. In words, a user stream contains all messages which are related with a certain set of users $u \in U'$ and all resources and further users which are related with these messages. On Twitter, examples of user stream aggregations include user lists and user directory streams. User list and user directory stream aggregation contain all messages which have been authored by a defined set of users and all resources and users related with these messages. While user list streams are maintained by the user who has created the list, user directory streams, such as the one provided by wefollow[4], allow users to add themselves to existing or new lists.

A message stream $S(M')$ is a tupel $S(M') = (U, M, R, Y', ft)$, where $Y' = \{(u, m, r) \,|\, m \in M'\}$ and $M' \subseteq M$ and $Y' \subseteq Y$. In words, a message stream contains all messages of a certain type (e.g. conversational direct messages or retweeted messages) and their related resources and users.

In addition all streams can be restricted to a specific time window in which the stream is recorded. For example, $S(M')[t_s, t_e]$ denotes a message stream recorded within the time window t_s and t_e. Formally $S[t_s, t_e]$ can be defined as follows: $S[t_s, t_e] = (U \times M \times R, Y, ft)$, where $ft : Y \to \mathbb{N}, t_s \leq ft \geq t_e$.

2.3 Properties of Social Awareness Streams

Since for a given keyword (e.g., *semantic web*) different types of social awareness stream aggregations (e.g., the *semanticweb* hashtag or keyword stream or various user directory or user list streams denoted by the label *semanticweb*) can be analyzed, we introduce several stream measures in order to be able to compare different stream aggregations and quantify their differences. It appears intuitive that different aggregations of social awareness streams would yield different stream properties and characteristics. However, as a community we know little about *how* our aggregation decisions influence what we can observe. For example: What kind of streams are most suitable to identify links to web resources or hashtags for a given user query? What kind of streams and what kind of network transformations are most suitable for identifying synonyms or hyponyms (e.g. for hashtags)? What kind of streams are effective for identifying experts for a given topic? What kind of streams are topically diverse vs. topically focussed and narrow?

In the following, we introduce a number of measures that can be applied to different aggregations of social awareness streams in order to answer such questions, and to enable a quantitative comparison of different stream aggregations.

2.3.1 *Social Diversity*

The social diversity of a stream measures the variety and balance of users authoring a stream, i.e. the social variety and social balance of a stream. The Stirling measure [20] captures three qualities of diversity: variety (i.e., how many individual users participate in a stream), balance (i.e., how evenly the participation is distributed among these users), and similarity (i.e., how related/similar those users are). That means, although we do not use the concepts of similarity yet, the proposed diversity measures could be extended by including the concept of similarity.

To measure the social variety we can count the number of unique users $|U_a|$ who authored messages in a stream. For normalization purposes we can include the stream size $|M|$. The social variety per message $SVpm$ represents the mean number of different authors per message and is defined as follows:

$$SVpm = \frac{|U_a|}{|M|} \tag{1}$$

The maximum social variety $SVpm$ of a social awareness stream is 1. A social variety $SVpm$ of 1 indicates that every message has been published by another user. The social variety can also be interpreted as a function which illustrates how the number of authors in a stream grows over time and with increasing number of messages. For example, the interpretation of the social variety over time is defined as follows:

$$SVpt(t) = \frac{|U_a[t]|}{|M[t]|} \tag{2}$$

The variable $|M[t]|$ represents the number of messages within the time interval t and $|U_a[t]|$ denotes the number of authors of these messages.

To quantify the social balance of a stream, we can define an entropy-based measures, which indicates how democratic a stream is. Specifically, we call the distribution of authors U_a for messages M of a given stream, $P(M|U_a)$. Given this

[4] `http://wefollow.com`

number, we define the social balance of a stream as follows:

$$SB = - \sum_{u \in U_a} P(m|u) * log(P(m|u)) \quad (3)$$

A low social balance indicates that a stream is dominated by few authors, i.e. the distribution of messages per author is not even. A high social balance indicates that the stream was created in a balanced way, i.e. the distribution of messages per author is even.

For example, if on a stream author A has published 3 messages, author B has published 1 message and author C has as well published 1 message in a stream, we would say the social balance of this stream is equal to:

$$SB = -\frac{3}{5} * log(\frac{3}{5}) - \frac{1}{5} * log(\frac{1}{5}) - \frac{1}{5} * log(\frac{1}{5}) \approx 1.37 \quad (4)$$

2.3.2 *Conversational Diversity*

The conversational diversity of a stream measures how many users communicate via a stream and can be approximated via the conversational variety and conversational balance of a stream. To measure the number of users being mentioned in a stream (e.g., via @replies or slashtags), we can introduce $|U_m|$ for $u_m \in U_m$. The conversational variety per message $CVpm$ represents the mean number of different users mentioned in one message of a stream and is defined as follows:

$$CVpm = \frac{|U_m|}{|M|} \quad (5)$$

The conversational variety can in the same way as the social variety be interpreted as a function over time and message. The conversational balance of a stream can be defined in the same way as the social balance, as an entropy-based measure (CB) which quantifies how predictable conversation participants are on a certain stream.

2.3.3 *Conversational Coverage*

From the number of conversational messages $|M_c|$ and the total number of messages of a stream $|M|$, we can compute the conversational coverage of a stream, which is defined as follows:

$$CC = \frac{|M_c|}{|M|} \quad (6)$$

The conversational coverage measures the mean number of messages of a stream that have a conversational purpose.

2.3.4 *Lexical Diversity*

The lexical diversity of a stream can be approximated via the lexical variety and lexical balance of a stream. To measure the vocabulary size of a stream, we can introduce $|R_k|$, which captures the number of unique keywords $r_k \in R_k$ in a stream. For normalization purposes, we can include the stream size ($|M|$). The lexical variety per message $LVpm$ represents the mean vocabulary size per message and is defined as follows:

$$LVpm = \frac{|R_k|}{|M|} \quad (7)$$

In the same way as the social variety we can interpret the lexical variety as a function which illustrates the growth of vocabulary over time and with increasing number of messages. The lexical balance LB of a stream can, in the same way as the social balance, be defined via an entropy-based measures which quantifies how predictable a keyword is on a certain stream.

2.3.5 *Topical Diversity*

The topical diversity of a stream can be approximated via the topical variety and topical balance of a stream. To compute the topical variety of a stream, we can use arbitrary surrogate measures for topics, such as the result of automatic topic detection or manual labeling methods. In the case of Twitter we could use the number of unique hashtags $r_h \in R_h$ as surrogate measure for topics. The topical variety per message $TVpm$ represents the mean number of topics per message and is defined as follows:

$$TVpm = \frac{|R_h|}{|M|} \quad (8)$$

The topical variety can also be interpreted as a function which illustrates the growth of the hashtag vocabulary over time and with increasing number of messages. The topical balance TB can, in the same way as the social balance, be defined as an entropy-based measures which quantifies how predictable a hashtag is on a certain stream.

2.3.6 *Informational Diversity*

The informational diversity of a stream can be approximated via the informational variety and informational balance of a stream. To measure the informational variety of a stream, we can compute the number of unique links in messages of a stream $|R_l|$ for $r_l \in R_l$. The informational variety per message $IVpm$ is defined as follows:

$$IVpm = \frac{|R_l|}{|M|} \quad (9)$$

In the same way as the social variety measure, the informational variety measure can be interpreted as a function which illustrates how the number of different links shared via a stream grows over time and with increasing number of messages. The informational balance IB can, in the same way as the social balance, be defined as an entropy-based measures which quantifies how predictable a link is on a certain stream.

2.3.7 *Informational Coverage*

From the number of informational messages $|M_i|$ and the total number of messages of a stream $|M|$ we can compute the informational coverage of a stream which is defined as follows:

$$IC = \frac{|M_i|}{|M|} \quad (10)$$

The informational coverage indicates how many messages of a stream have a informational character.

2.3.8 *Spatial Diversity*

The spatial diversity of a stream measures the variety and balance of geographical message annotations in a stream, i.e. the spatial variety and spatial balance of a stream. The more spatial diverse a stream is the more messages it contains which were published on different locations and the more even the message distribution is across these locations. The spatial variety per message $SPVpm$ of a stream is defined via the number of unique locations of messages in a stream $|L|$

and the number of messages $|M|$ and is defined as follows:

$$SPVpm = \frac{|L|}{|M|} \tag{11}$$

In the same way as the social variety measure, the spatial variety measure can be interpreted as a function which illustrates how the number of different geo-locations grows over time and with increasing number of messages. The spatial balance SPB can, in the same way as the social balance, be defined as an entropy-based measures which quantifies how balanced messages are distributed across these geo-locations.

2.3.9 Temporal Diversity

The temporal diversity of a stream can be approximated via the temporal variety and temporal balance of a stream. The more temporal diverse a stream is the more messages it contains which were published at different moment in time and the more even the message distribution is across these timestamps. The temporal variety per message $TPVpm$ of a stream is defined via the number of unique timestamps of messages $|TP|$ and the number of messages $|M|$ in a stream and is defined as follows:

$$TPVpm = \frac{|TP|}{|M|} \tag{12}$$

In the same way as the social variety, the temporal variety measure can be interpreted as a function which illustrates how the number of different timestamps grows over time and with increasing number of messages. The temporal balance TPB can, in the same way as the social balance, be defined as an entropy-based measures which quantifies how balanced messages are distributed across these message-publication-timestamps.

3. METHODOLOGY AND EXPERIMENTAL SETUP

To explore the nature of different social awareness stream aggregations which can be created for a given keyword and semantic models emerging from them, we conducted the following experiments. We studied different social awareness streams for the topic *semantic web* which were all recorded within the same time window. We investigated stream properties and semantics by adopting the introduced measures and by applying various network transformations.

Since measures for similarity and relatedness are not well developed for three-mode networks yet, the tripartite structure is often reduced to 3 two-mode networks with regular edges. These 3 networks model the relations between resources and users (N_{RU} network), resources and messages (N_{RM} network) and messages and users (N_{MU} network). To avoid subsubscriptions from now on we use RM, RU and MU instead of N_{RM}, N_{RU} and N_{MU}.

For example, the resource-user network RU can be defined as follows: $RU = (R \times U, E_{ru}), E_{ru} = \{(r,u) \mid \exists i \in I : (r,u,i) \in E\}, w : E \rightarrow \mathbb{N}, \forall e = (r,u) \in E_{ru}, w(e) := |i : (r,u,i) \in E|$. In words, the two-mode network RU links users to the resources that they have used or with which they have been mentioned in at least one message. Each link is weighted by the number of times a user has used or has been mentioned with that resource. The RU network can be represented as a matrix of the form $RU = v_{ij}$ where $v_{ij} = 1$ if user u_i is related with resource r_j. Since the resource-user network RU is an unqualified network, several qualified or semi-qualified networks (e.g. the resource-author network RU_a or the hashtag-author network R_hU_a), which are specializations of the resource-user network, can be deduced.

The resource-message RM and message-user MU networks are defined in the same way as the resource-user network: In words, the two-mode network RM links resources to messages in which they have been used at least once. Each link is weighted by the number of times a resource was used in a message. The two-mode network MU links messages to users which have authored them or are mentioned in them. Each link is weighted by the number of times a message was related with a user.

In order reveal associations between resources, we extracted non-qualified resource-message networks RM and semi-qualified resource-author networks RU_a from different social awareness stream aggregations. By multiplying the corresponding two-mode network matrices with their transpose (e.g., $O_R(RM) = RM * RM^T$), we transformed them into non-qualified one-mode networks of resources ($O_R(RM)$ and $O_R(RU_a)$), which can be considered as lightweight, associative resource ontologies [16]. From these non-qualified resource ontologies, we extracted semi-qualified resource networks, namely resource-hashtag networks RR_h and resource-link network RR_l, which we again transformed into associative resource ontologies ($O_R(RR_h(RM))$, $O_R(RR_l(RM))$, $O_R(RR_h(RU_a))$ and $O_R(RR_l(RU_a))$). Different ontologies relate resources which occur in the same contexts of messages/users/hashtags/links and therefore tend to have similar meanings according to Harris' distributional hypothesis [2].

The qualities of different resource ontologies depend on the different ways they are created: For example the $O_R(RM)$ ontology relates resources which co-occur in different messages and weight their relations by the number of times they co-occur. That means, a strong association exists between two resources if they share a large percentage of messages, regardless whether these associations were created by the same users or not. The $O_R(RU_a)$ ontology relates resources which are used by the same users. Relations between resources are weighted by the number of individual users who have used both resources, regardless whether these resources were used in one or different messages of them. The $O_R(RR_h(RM))$ and $O_R(RR_h(RU_a))$ network weight relations between resources by the number of times they co-occur with common hashtags. That means, a strong association exists between two resources if they share a large percentage of hashtags. The $O_R(RR_l(RM))$ and $O_R(RR_l(RU_a))$ network weights relations between resources by the number of times they co-occur with common URLs. That means, between two resources exists a strong association if they share a large percentage of links. In the $O_R(RR_l(RM))$ and $O_R(RR_h(RM))$ network resources cooccur if they are related with the same message (regardless whether these resources were associated via one or several users), while in the $O_R(RR_l(RU_a))$ and $O_R(RR_h(RU_a))$ network resources co-occur if they have been authored by the same user (regardless whether these resources were used in one or several messages of one user).

Since the different qualities of resource ontologies heavily depend on the different two-mode networks from which they originate, we also compared different two-mode networks in terms of their most important resource rankings. As a reminder, in the resource-message network RM a resource is important if it occurs in many different messages, while in the resource-author network RU_a a resource is important if it is used by many different users. In the resource-hashtag networks, $RRh(RM)$ and $RRh(RU_a)$, a resource is important if it co-occurs with many different hashtags. In the resource-link networks, $RRl(RM)$ and $RRl(RU_a)$, the resource ranking depends on the number of different links with which a resource co-occurs. If for example a resource #semanticweb appears in 50 percent of all messages of a stream which have all been generated by one user, this resource would have a high rank in the resource-message RM network, but a very low rank in the resource-author RU_a network. If the resource #semanticweb occurs together with certain URL in a message, which was retweeted many times by different users, the resource #semanticweb would have a high rank in the resource-message RM and resource-author RU_a network, but a very low rank in the resource-link RR_l and resource-hashtag RR_h network.

To assess the quality of different two-mode networks we assumed that hashtags tend to be semantic richer than other resources, because hashtags are often used to add additional contextual information to messages. Under this assumption we were able to quantitatively assess the semantic richness of different two-mode networks by computing the number of hashtags which appear under the top n resources (for n=15, 50, 100).

3.1 Dataset

We analyzed and compared the following social awareness stream aggregations from Twitter which were all related to one topic, *semantic web*. The stream aggregations were recorded in 2 time intervals: from 16th of Dec 2009 to 20th of Dec 2009 and from 29th of Dec 2009 to 1st of Jan 2010. While the first time interval represents 4 "normal" days without specific events or disturbances, we included the second time window due to the occurrence of a particular event (New Years Day) to surface differences in different stream aggregations.

- The *semanticweb* hashtag stream[5] $S(R_h)$ is a resource stream which includes all public messages containing the resource #semanticweb and all resources and users related with these messages. $S(R_h)$ is defined as follows: $S(R_h) = (U, M, R, Y', ft)$, where $Y' = \{(u, m, r) \,|\, r \in \{\#semanticweb\} \vee \exists r' \in \{\#semanticweb\}, \tilde{m} \in M, u \in U : (u, \tilde{m}, r') \in Y\}$ where $R_h \subseteq R$ and $Y' \subseteq Y$.

- The *semanticweb* keyword stream[6] $S(R_k)$ consists of all public messages containing the keyword semanticweb and semweb, a common abbreviation, and all resources and users related with these messages. $S(R_k)$ is defined as follows: $S(R_k) = (U, M, R, Y', ft)$, where $Y' = \{(u, m, r) \,|\, r \in \{semanticweb, semweb\} \vee \exists r' \in \{semanticweb, semweb\}, \tilde{m} \in M, u \in U : (u, \tilde{m}, r') \in Y\}$ where $R_k \subseteq R$ and $Y' \subseteq Y$.

- The *semweb* user list stream[7] $S(U_{UL})$ is a user stream which contains all public messages published by users of the authoritatively defined *semweb* user list and all resources and users related with these messages. We have chosen this list, because of its high authority for the topic *semantic web*. The list was created by Stefano Bertolo (*user sclopit*[8]), who is a Project Officer at the European Commission in the field of Knowledge Representation and Content Management. At the time we crawled the list (23th of November 2009), 141 users $u \in U_{UL}$ were included. $S(U_{UL})$ is defined as follows: $S(U_{UL}) = (U, M, R, Y', ft)$, where $Y' = \{(u, m, r) \,|\, u \in U_{UL} \vee u' \in U_{UL}, \tilde{m} \in M, r \in R : (u', \tilde{m}, r) \in Y\}$ where $U_{UL} \subseteq U$ and $Y' \subseteq Y$.

Stream	$\|M\|$	$\|U_a\|$	$\|U_m\|$	$\|R_k\|$	$\|R_h\|$	$\|R_l\|$
$S(R_h)$	156	60	41	182	103	111
$S(R_k)$	210	105	66	618	108	133
$S(U_{UL})$	2183	86	770	4683	544	898
$S(U_{UD})$	4544	139	1559	6059	805	1300

Table 1: Number of messages ($|M|$), authors ($|U_a|$), users ($|U_m|$), keywords($|R_k|$), hashtags ($|R_h|$), and links ($|R_l|$) mentioned in messages of hashtag $S(R_h)$, keyword $S(R_k)$, user list $S(U_{UL})$, and user directory $S(U_{UD})$ stream aggregations.

- The *semanticweb* wefollow user directory stream[9] $S(U_{UD})$ is a user stream which contains all public messages of users of the collaboratively created *semanticweb* directory and all resources and users related with these messages. We have chosen this directory, because it contains a large number of users. At the time we crawled the directory (23th of November 2009) it consisted of 191 users $u \in U_{UD}$. $S(U_{UD})$ is defined as follows: $S(U_{UD}) = (U, M, R, Y', ft)$, where $Y' = \{(u, m, r) \,|\, u \in U_{UD} \vee u' \in U_{UD}, \tilde{m} \in M, r \in R : (u', \tilde{m}, r) \in Y\}$ where $U_{UD} \subseteq U$ and $Y' \subseteq Y$.

3.2 Properties of Different Twitter Streams

To analyze and compare different stream aggregations we computed serval basic stream properties (see Table 1) and previously defined stream measures (see Figure 2).

From Figure 2 we can see that both analyzed resource streams (i.e., the hashtag and keyword stream) have a slightly higher informational variety *IVpm* and informational coverage *IC* than the analyzed user streams (i.e., user list and user directory streams). This result suggests that researchers who want to sample messages from social awareness streams that contain links would benefit from focusing on hashtag or keyword streams (as opposed to other types of streams).

Figure 2 also shows that both analyzed resource streams have a higher social diversity (which is reflected via the social variety (*SVpm*) and social balance (*SB*) measure) than the analyzed user streams. Specially, if we compare the social balance (*SB*) of different stream aggregations, we can see that the analyzed hashtag stream has a significant higher social balance. This indicates that hashtag streams may be more democratic than other types of streams, since the participation of different authors (i.e., the number of messages they produce) seems to be more balanced.

[5] http://twitter.com/search?q=\%23semanticweb

[6] http://twitter.com/\#search?q=semanticweb

[7] http://twitter.com/sclopit/semweb

[8] http://twitter.com/sclopit

[9] http://wefollow.com/twitter/semanticweb

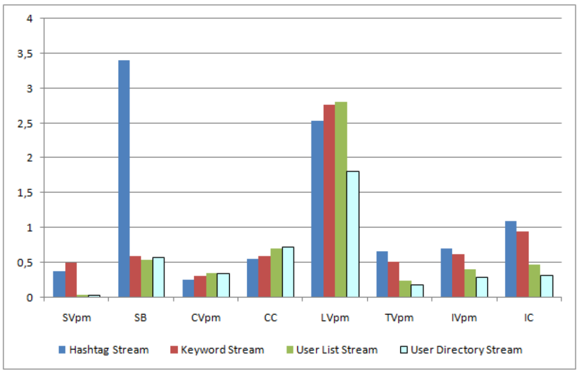


Figure 2: Social- (SVpm), Conversational- (CVpm), Lexical- (LVpm), Topical- (TVpm) and Informational (IVpm) Variety per message, Social Balance (SB), Informational- (IC) and Conversational Coverage (CC) of different Social Awareness Streams.

Since user list streams are closed and authoritatively defined sets of users, it seems plausible that these streams would contain less participants compared to open resource streams. However, the fact that the number of messages contained in the user directory stream is more than double the messages contained in the user list stream (although the number of authors is less than 40 percent higher) indicates that users registered in a user directories produce more messages. Since the social balance of the user directory stream is rather low, few authors seem to produce a major part of messages.

It is also interesting to note that the topical variety $TVpm$ is higher for the analyzed resource streams as for the analyzed user streams. Figure 2 shows that in a hashtag stream, more than every second message contains another hashtag (in addition to the one hashtag which is needed to assign the message to the hashtag stream), whereas the hashtag quota of other streams is lower.

3.3 Results

The aim of our empirical work was to explore if and what kind of knowledge can be acquired from different aggregations of social awareness streams by transforming them into lightweight, associative resource ontologies. The lightweight ontologies expose how related two resources are but do not contain any information about the semantics of relations.

In the following, we present our first results of analyzing emerging semantics conveyed by one user stream (the *semweb* user list stream $S(U_{UL})$) and one resource stream (the *#semanticweb* hashtag stream $S(R_h)$), which are both related with the topic *semantic web*.

Table 2 gives qualitative insights into the emerging semantics of different two-mode networks which were later transformed into resource ontologies. From Table 2 we can see 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" (see Figure 3).

If we compare the 15 most important resources in different networks extracted from the same authoritative list of users (the *semweb* user list $S(U_{UL})$), we can observe that in all of them (except in one) the most important resources are mainly words which are not relevant for the topic *semantic web*. Only the resource-hashtag network $RR_h(RM)S(U_{UL})$ seems to be a positive exception and ranks resources (such as `#linkeddata, data, #goodrelations, #semanticweb, source, #distributed, link, #http, #rdf, page, great, web`) high, which are obviously relevant for the topic *semantic web*. This indicates that in a user stream of experts for a certain topic, resources which co-occur with many different hashtags tend to be very relevant for the expertise topic (or topic of common interest) of the group. A more detailed look into the most frequent hashtags of the analyzed user list stream (e.g., `#linkeddata, #semanticweb, #googrelations, #rdf, #rdfa`) confirms this assumption. One possible explanation for this phenomenon is that experts use a very fine-granular vocabulary to talk about their expertise topic and create a detailed hashtag vocabulary to add additional information to their messages and to assign them to appropriate communication channels.

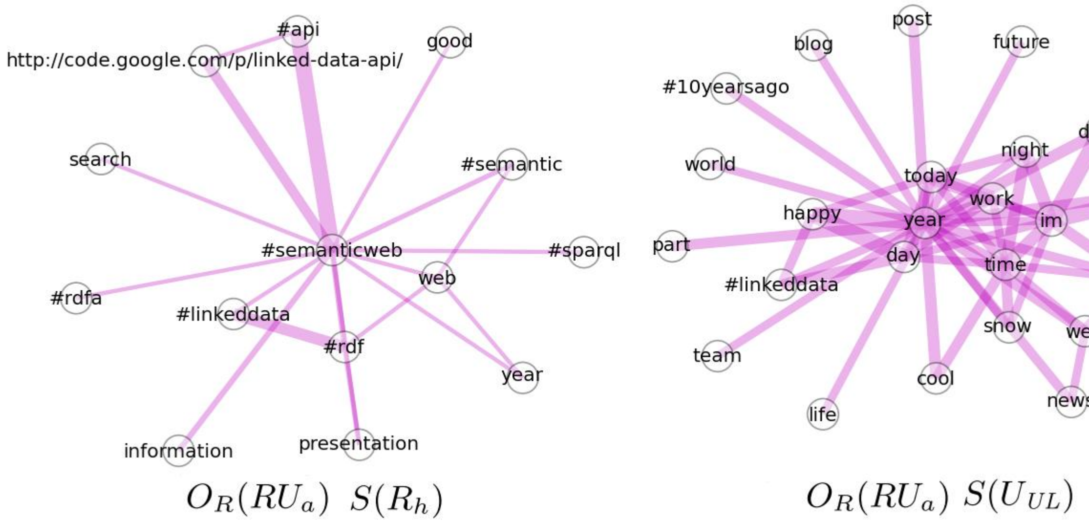
The good quality of the $O_R(RR_h(RM))S(U_{UL})$ ontology, compared to other ontologies extracted from the user list stream aggregation, can amongst others be explained through hashtags' quality of revealing contextual information. Hashtags seem to be more appropriate for estimating the context of resources and identifying semantic similar resources via their common contexts.

For us it was surprising that URLs do not show similar characteristics as hashtags. At the beginning of our work we assumed that URLs might be as well a very appropriate context indicator. However, resource ontologies generated from resource-link networks do not reveal relevant concepts for the topic *semantic web*. These ontologies contain many general resources such as `type`, `source`, `blog` and `read`. These resources heavily occur with many common links, but do not reveal interesting knowledge about the stream aggregation topic.

4. DISCUSSION OF RESULTS

Our empirical findings confirmed our assumption 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". Nevertheless, it would be reasonable to assume that a stream of messages produced by experts in a given domain would result in meaningful semantic models describing resources within the domain and relations between them. Our findings however suggest that this is not necessarily the case. Not only are user list streams prone to external disturbances, the different types of network transformations also influence the resulting semantics.

Research on emerging semantics from folksonomies [16] showed that ontologies extracted from concept-instance networks (which are equivalent to resource-message networks in our model) are more appropriate for concept-mining than concept-user networks (which are equivalent to resource-user networks in our model), but ignore the relevance of individual concepts from the user perspective. Therefore, concept-instance networks might give an inaccurate picture of the community. This line of research would suggest to compute resource ontologies from resource-user networks rather than resource-message networks of social awareness stream aggregations in order to get an accurate picture of the community participating in the stream. Our results however indicate that resource-author networks (and ontologies gen-



$O_R(RU_a)\ S(R_h)$ $O_R(RU_a)\ S(U_{UL})$

Figure 3: The resource ontology $O_R(RU_a)S(R_h)$ computed from the resource-user network of the `#semanticweb` hashtag stream shows an emerging semantic model which is able to describe the meaning of the stream's label *#semanticweb*, while the $O_R(RU_a)S(U_{UL})$ ontology computed from the resource-user network of the *semweb* user list stream shows that resource-user network transformations are perceptible for disturbances.

erated from them) are very prone to "disturbances", such as New Years Eve or the Avatar movie start, because these networks relate resources if they have common authors (regardless if they were used in one or several messages). Therefore, if for example all users post one message which contains happy new year greetings, resources such as `happy` or `year` become very important, although the majority of messages in this stream might be about *semantic web*. Our results indicate that hashtag-resource transformations have the power to reduce the non-informational noise of social awareness streams and reveal meaningful semantic models describing the domain denoted by the stream aggregation label (e.g., *semantic web*).

5. RELATED WORK

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

Our work was inspired by Mika's work [16] who explored different lightweight, associative ontologies which emerge from folksonomies through simple network transformations. In general, automatic construction of term hierarchies and ontologies has been studied in both, the information retrieval and the semantic web communities: Sanderson and Croft describe in [18] the extraction of concept hierarchies from a document corpus. They use a simple statistical model for subsumption and apply it to concept terms extracted from documents returned for a directed query. Another line of research (e.g., [3]) suggests to use lexico-syntactic patterns (e.g., "such as") to detect hyponymy relations in text. Finally, the use of hierarchical clustering algorithms for automatically deriving term hierarchies from text was, amongst others, proposed in [1].

Since on the Social Web new data structures such as folksonomies (consisting of users, tags and resources) emerge, the extension and adaption of traditional content and link analysis algorithms and ontology learning algorithm became a key question. Markines et al. [15] define, analyze and evaluate different semantic similarity relationships obtained from mining socially annotated data. Schmitz et al. [19] describe how they mine from a tag space association rules of the form *If users assign the tags from X to some resource, they often also assign the tags from Y to them.* If resources tagged with t_0 are often also tagged with t_1 but a large number of resources tagged with t_1 are not tagged with t_0, t_1 can be considered to subsume t_0. Mika [16] presents a graph-based approach and shows how lightweight ontologies can emerge from folksonomies in social tagging systems. For mining concept hierarchies he adopts the set-theoretic approach that corresponds to mining association rules as is described by Schmitz et al.. Heymann at al. [4] represents each tag t as a vector (of resources tagged with the tag) and computes cosine similarity between these vectors. That means, they compute how similar the distributions of tags are over all resources. To create a taxonomy of tags, they sort the tags according to their closeness-centrality in the similarity graph. They start with an empty taxonomy and add a tag to the taxonomy as a child of the tag it is most similar to, or as a root node if the similarities are below a threshold.

In our past work we studied quantitative measures for tagging motivation [12] and found empirical evidence that the emergent semantics of tags in folksonomies are influenced by the pragmatics of tagging, i. e. the tagging practices of individual users [11]. This work as well was inspired by the hypothesis that the quality of emergent semantics (what concepts mean) depends on the pragmatics of users participating in a stream (how concepts are used).

	Top 15 resources
$RM\ S(R_h)$	#semanticweb, semantic, source, web, #linkeddata, twitter, #rdf, data, link, 2010, technology, present, #singularity, tool, #ontology
$RU_aS(R_h)$	#semanticweb, semantic, web, #rdf, #linkeddata, data, link, present, #sparql, good, 2010, technology, search, #semantic, http://code.google.com/p/linked-data-api/
$RR_h(RM)S(R_h)$	#semanticweb, source, #linkeddata, semantic, data, link, web, #rdf, 2010, twitter, present, http://ouseful.wordpress.com/2009/12/15, pipelink, http://code.google.com/p/linked-data-api/ , #sparql
$RR_h(RU_a)S(R_h)$	#semanticweb, semantic, web, #linkeddata, #rdf, data, link, http://code.google.com/p/linked-data-api/ , #semantic, #api, present, people, nobot, real, explain
$RR_l(RM)S(R_h)$	#semanticweb, source, semantic, twitter, web, #linkeddata, #rdf, tool, present, link, data, technology, #singularity, 800, entry
$RR_l(RU_a)S(R_h)$	#semanticweb, web, semantic, #rdf, tool, #linkeddata, 800, entry, exce, technology, list, source, #wiki, link, #owl
$RM\ S(U_{UL})$	type, year, data, good, #linkeddata, time, imo, 2010, source, great, make, web, work, day, watch
$RU_aS(U_{UL})$	year, make, great, 2010, work, web, day dont, happy, good, time, imo, interest, data, nice
$RR_h(RM)S(U_{UL})$	#linkeddata, data, #goodrelations, #semanticweb, source, #distributed, link, #http, #rdf, page, great, web, good, #bold, work
$RR_h(RU_a)S(U_{UL})$	year, make, happy, data, day, 2010, web, dont, great, interest, time, today, page, idea, future
$RR_l(RM)S(U_{UL})$	type, source, #semanticweb, #linkeddata, 2010, data, web, semantic, blog, state, post, make, new, twitter, read
$RR_l(RU_a)S(U_{UL})$	make, work, people, cool, time, read, thing, blog, new, book, help, language, change, talk, post

Table 2: Most important resources (ranked via their frequency) extracted from the resource-message (RM), the resource-author (RU_a), the resource-hashtag ($RR_h(RM)$ and $RR_h(RU_a)$), and the resource-link ($RR_l(RM)$ and $RR_l(RU_a)$) networks of a selected hashtag stream $S(R_h)$ and user list stream $S(U_{UL})$

Our work differs from existing work (1) through our focus on social awareness streams which have a more complex and dynamic structure than folksonomies and (2) through our focus on stream aggregations and data preprocessing. The aim of this work was to explore the initial step of building ontologies from social awareness streams, i.e. to explore how different stream aggregation and simple network transformations can influence what we can observe.

In general, little research on social awareness streams exists to date. Some recent research investigates user‘s motivation for microblogging and microblogging usage by analyzing user profiles, social interactions and activities on Twitter: A study by [8] shows that the rate of user activities on Twitter is driven by the social network of his actual friends. Users with many friends tend to post more updates than users with few friends. The work distinguishes between two different social networks of a user, the “declared” social network made up of followers and followees and the sparser and simpler network of actual friends. In [13], the authors performed a descriptive analysis of the Twitter network. Their results indicate that frequent updates might be correlated with high overlap between friends and followers. The work of [10] provides many descriptive statistics about Twitter use, and hypothesizes that the differences between users network connection structures can be explained by three types of distinct user activities: information seeking, information sharing, and social activity. In [21] an algorithm for identifying influential Twitter users for a certain topic is presented.

Other research focuses on analyzing content of social awareness stream messages, e.g. to categorize or cluster them or to explore conversations. For example, in [6] the authors examined the functions and usage of the @ (“reply/mention”) symbol on Twitter and the coherence of conversations on Twitter. Using content analysis, this line of work developed a categorization of the functional use of @ symbols, and analyzed the content of the reply messages. Recent research explores sentiments, opinions and comments about brands exposed on Twitter [9] and produces characterization of the content of messages of social awareness streams [17]. Naaman et al. examine how message content varies by user characteristics, personal networks, and usage patterns.

In the light of existing research and to the best of our knowledge, the network-theoretic model introduced in our paper represents the first attempt towards formalizing different aggregations of social awareness streams.

6. CONCLUSION AND FUTURE WORK

As our knowledge about the nature and properties of social awareness streams is still immature, this paper aimed to make following contributions: 1) We have introduce a network-theoretic model of social awareness streams, a so-called tweetonomy, which provides a formal, extensible framework capable of accommodating the complex and dynamic structure of message streams found in applications such as Twitter or Facebook. 2) We have defined and applied a number of measures to capture interesting characteristics and properties of different aggregations of social awareness streams and 3) Our empirical work shows that different aggregations of social awareness streams exhibit interesting different semantics.

While the network-theoretic model of social awareness streams is general, the empirical results of this paper are limited to a single concept (*semantic web*). It would be interesting to expand our analysis to a broader variety of social awareness streams and to conduct experiments over greater periods of time. For example, it seems plausible to assume that streams for hashtags such as `#www2010` or `#fun` would differ significantly from a stream for the hashtag `#semanticweb`. We leave the task of applying our model to the analysis of a broader set of social awareness streams to future research. When it comes to the semantic analysis of social awareness streams, the extent to which different streams approximate the semantic understanding of users that are participating in these streams is interesting to investigate. While we have tackled this issue by selecting a narrow domain (*semantic web*), more detailed evaluations that include user feedback are conceivable. In addition, the semantic analysis conduced is based on simple network transformations. In future work, it would be interesting to study whether more sophisticated knowledge acquisition methods which, for example, exploit external background knowledge (such as WordNet[10] and DBpedia[11]) would produce different results. Another interesting issue raised by our investigations is the extent to which the semantics of social awareness streams are influenced by tweeting pragmatics of individual users or user groups and vice versa.

The network-theoretic model of this paper is relevant for researchers interested in information retrieval and ontology learning from social awareness streams. The introduced stream measures are capable of identifying interesting differences and properties of social awarness streams. Our empirical results provide evidence that *there is some semantic "wisdom" in aggregated streams of tweets*, but different stream aggregations exhibit different semantics and different extraction methods influence resulting semantic models: While some semantic models and aggregations of streams are rather robust against external events (such as New Years Day), other models and aggregations of streams are more perceptible to such "disturbances", and lend themselves to different purposes.

6.1 Acknowledgments

The research presented in this work is in part funded by the Know-Center and the FWF Austrian Science Fund Grant P20269 TransAgere. The authors want to thank Jan Pöschko and Florian Klien for their valuable comments on this work and Ronald Fellmann for supporting us with data aggregation issues.

7. REFERENCES

[1] P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research (JAIR)*, 24:305–339, 2005.

[2] Z. Harris. Distributional structure. *The Structure of Language: Readings in the philosophy of language*, 10:146–162, 1954.

[3] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 539–545, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

[4] P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Computer Science Department, April 2006.

[5] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 195–206, New York, NY, USA, 2008.

[6] C. Honey and S. C. Herring. Beyond microblogging: Conversation and collaboration via twitter. In *System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on System Sciences*, 2009.

[7] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Bibsonomy: A social bookmark and publication sharing system. In *Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures*, pages 87–102, 2006.

[8] B. A. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *ArXiv e-prints*, December 2008.

[9] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2169–2188, 2009.

[10] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, New York, NY, USA, 2007.

[11] C. Körner, D. Benz, A. Hotho, M. Strohmaier, and G. Stumme. Stop thinking, start tagging: Tag semantics emerge from collaborative verbosity. In *19th International World Wide Web Conference (WWW2010)*. ACM, April 2010.

[12] C. Körner, R. Kern, H. Grahsl, and M. Strohmaier. Of categorizers and describers: An evaluation of quantitative measures for tagging motivation. In *21st ACM SIGWEB Conference on Hypertext and Hypermedia (HT2010)*. ACM, June 2010.

[13] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 19–24, New York, NY, USA, 2008.

[14] R. Lambiotte and M. Ausloos. Collaborative tagging as a tripartite network, Dec 2005.

[15] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 641–650, New York, NY, USA, 2009.

[16] P. Mika. Ontologies are us: A unified model of social networks and semantics. *Web Semant.*, 5(1):5–15, 2007.

[17] M. Naaman, J. Boase, and C.-H. Lai. Is it all about me? user content in social awareness streams. In *Proceedings of the ACM 2010 conference on Computer supported cooperative work*, 2010.

[18] M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–213, New York, NY, USA, 1999. ACM.

[19] P. Schmitz. Inducing ontology from flickr tags. In *Proceedings of the Workshop on Collaborative Tagging at WWW2006*, Edinburgh, Scotland, May 2006.

[20] A. Stirling. A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, 4(15):707–719, August 2007.

[21] J. Weng, E. peng Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *Third ACM International Conference on Web Search and Data Mining (WSDM 2010)*, 2010.

[10] `http://wordnet.princeton.edu/`

[11] `http://dbpedia.org/`