Semantic Web 0 (0) 1
IOS Press

Making Sense of Social Media Streams through Semantics: a Survey

Editor(s): Andreas Hotho, University of Würzburg, Germany
Solicited review(s): Harald Sack, Universität Potsdam, Germany; Ashutosh Jadhav, Wright State University, USA
Open review(s): Anonymous Reviewer

Kalina Bontcheva [a,*] Dominic Rout [a]
[a] *Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello, Sheffield, United Kingdom*
E-mail: Initial.Surname@dcs.shef.ac.uk

Abstract.
Using semantic technologies for mining and intelligent information access to social media is a challenging, emerging research area. Traditional search methods are no longer able to address the more complex information seeking behaviour in media streams, which has evolved towards sense making, learning, investigation, and social search. Unlike carefully authored news text and longer web context, social media streams pose a number of new challenges, due to their large-scale, short, noisy, context-dependent, and dynamic nature.

This paper defines five key research questions in this new application area, examined through a survey of state-of-the-art approaches for mining semantics from social media streams; user, network, and behaviour modelling; and intelligent, semantic-based information access. The survey includes key methods not just from the Semantic Web research field, but also from the related areas of natural language processing and user modelling. In conclusion, key outstanding challenges are discussed and new directions for research are proposed.

Keywords: semantic annotation, semantic-based user modelling, semantic search, information visualisation, social media streams

1. Introduction

The widespread adoption of social media is based on tapping into the social nature of human interactions, by making it possible for people to voice their opinion, become part of a virtual community and collaborate remotely. If we take micro-blogging as an example, Twitter has 100 million active users, posting over 230 million tweets a day[1].

Engaging actively with such high-value, high-volume, brief life-span media streams has now become a daily challenge for both organisations and ordinary people. Automating this process through intelligent, semantic-based information access methods is therefore increasingly needed. This is an emerging research area, combining methods from many fields, in addition to semantic technologies, namely natural language processing, social science, machine learning, personalisation, and information retrieval.

Traditional search methods are no longer able to address the more complex information seeking behaviour in social media, which has evolved towards sense making, learning and investigation, and social search [107]. Semantic technologies have the potential to help people cope better with social media-induced information overload. Automatic semantic-based methods that adapt to individual's information seeking goals and summarise briefly the relevant social media, could ultimately support information interpretation and decision making over large-scale, dynamic media streams.

*Corresponding author. E-mail: K.Bontcheva@dcs.shef.ac.uk.
[1] http://www.guardian.co.uk/technology/pda/2011/sep/08/twitter-active-users (Visited May 7, 2012)

1570-0844/0-1900/$27.50 © 0 – IOS Press and the authors. All rights reserved

Unlike carefully authored news and other textual web content, social media streams pose a number of new challenges for semantic technologies, due to their large-scale, noisy, irregular, and social nature. In this paper we discuss the following key research questions, examined through a survey of state-of-the-art approaches:

1. What ontologies and Web of Data resources can be used to represent and reason about the semantics of social media streams?
2. How can semantic annotation methods capture the rich semantics implicit in social media?
3. How can we extract reliable information from these noisy, dynamic content streams?
4. How can we model users' digital identity and social media activities?
5. What semantic-based information access methods can help address the complex information seeking behaviour in social media?

To the best of our knowledge, this is the first comprehensive meta-review of semantic technology for mining and intelligent information access, where the focus is on current limitations and outstanding challenges, specifically arising in the context of social media streams.

The paper is structured as follows: section 2 provides background on social media, their different characteristics, and the corresponding technological challenges. Section 3 focuses on ontologies which model different kinds of social media, user profiles and networks, information sharing, and other typical social media activities (research question 1). Section 4 discusses methods for semantic annotation of social media streams, in particular the ways in which they capture the rich implicit semantics (research question 2) and deal with the noisy, streaming nature of this type of content (research question 3). Section 5 investigates in depth research question 4, i.e. how are users, networks, and activities modelled semantically and how can this knowledge be used to personalise information access. Next, section 6 analyses state-of-the-art in intelligent information access for social media streams, in the context of research question 5. In conclusion, section 7 defines outstanding challenges and provides directions for future work.

2. Social Media Streams: Characteristics, Challenges and Opportunities

Social media sites allow users to connect with each other for the purpose of sharing content (e.g. web links, photos, videos), experiences, professional information, and online socialising with friends. Users create posts or status updates and social media sites circulate these to the user's social network. The key difference from traditional web pages is that users are not just passive information consumers, but many are also prolific content creators.

Social media can be categorised on a spectrum, based on the type of connection between users, how the information is shared, and how users interact with the media streams:

- Interest-graph media [115], such as Twitter, encourage users to form connections with others based on shared interests, regardless of whether they know the other person in real life. Connections do not always need to be reciprocated. Shared information comes in the form of a stream of messages in reverse chronological order.
- Social networking sites (SNS) encourage users to connect with people they have real-life relationships with. Facebook, for example, provides a way for people to share information, as well as comment on each other's posts. Typically, short contributions are shared, outlining current events in users' lives or linking to something on the internet that users think their friends might enjoy. These status updates are combined into a time-ordered stream for each user to read.
- Professional Networking Services (PNS), such as LinkedIn, aim to provide an introductions service in the context of work, where connecting to a person implies that you vouch for that person to a certain extent, and would recommend them as a work contact for others. Typically, professional information is shared and PNS tend to attract older professionals [130].
- Content sharing and discussion services, such as blogs, video sharing (e.g. YouTube, Vimeo), slide sharing (e.g. SlideShare), and user discussion/review forums (e.g. CNET). Blogs usually contain longer contributions. Readers might comment on these contributions, and some blog sites create a time stream of blog articles for followers to read. Many blog sites also advertise automatically new blog posts through their users' Facebook and Twitter accounts.

These different kinds of social media, coupled with their complex characteristics, make semantic interpretation extremely challenging. State-of-the-art automatic semantic annotation, browsing, and search algorithms have been developed primarily on news articles and other carefully written, long web content [20]. In contrast, most social media streams (e.g. tweets, Facebook messages) are strongly inter-connected, temporal, noisy, short, and full of slang, leading to severely degraded results[2].

These challenging social media characteristics are also opportunities for the development of new semantic technology approaches, which are better suited to media streams:

Short messages (microtexts): Twitter and most Facebook messages are very short (140 characters for tweets). Many semantic-based methods reviewed below supplement these with extra information and context coming from embedded URLs and hashtags[3]. For instance, Abel *et al* [2] augment tweets by linking them to contemporaneous news articles, whereas Mendes *et al* exploit online hashtag glossaries to augment tweets [87].

Noisy content: social media content often has unusual spelling (e.g. 2moro), irregular capitalisation (e.g. all capital or all lowercase letters), emoticons (e.g. :-P), and idiosyncratic abbreviations (e.g. ROFL, ZOMG). Spelling and capitalisation normalisation methods have been developed [57], coupled with studies of location-based linguistic variations in shortening styles in microtexts [53]. Emoticons are used as strong sentiment indicators in opinion mining algorithms (see Section 4.4).

Temporal: in addition to linguistic analysis, social media content lends itself to analysis along temporal lines, which is a relatively under-researched problem. Addressing the temporal dimension of social media is a pre-requisite for much-needed models of conflicting and consensual information, as well as for modelling change in user interests. Moreover, temporal modelling can be combined with opinion mining, to examine the volatility of attitudes towards topics over time.

Social context is crucial for the correct interpretation of social media content. Semantic-based methods need to make use of social context (e.g. who is the user connected to, how frequently they interact), in order to derive automatically semantic models of social networks, measure user authority, cluster similar users into groups, as well as model trust and strength of connection.

User-generated: since users produce, as well as consume social media content, there is a rich source of explicit and implicit information about the user, e.g. demographics (gender, location, age, etc.), interests, opinions. The challenge here is that in some cases, user-generated content is relatively small, so corpus-based statistical methods cannot be applied successfully.

Multilingual: Social media content is strongly multilingual. For instance, less than 50% of tweets are in English, with Japanese, Spanish, Portuguese, and German also featuring prominently [26]. Unfortunately, semantic technology methods have so far mostly focused on English, while low-overhead adaptation to new languages still remains an open issue. Automatic language identification [26,10] is an important first step, allowing applications to first separate social media in language clusters, which can then be processed using different algorithms.

The rest of this paper discusses which of these challenges have been addressed by semantic technologies to date and how.

3. Ontologies for Representing Social Media Semantics

Ontologies are the corner stone of semantic technology applications. In this section we focus specifically on ontologies created to model different kinds of social media, user profiles, sharing, tagging, liking, and other common user behaviour in social media. Table 1 provides an overview of these ontologies, alongside different dimensions, which are discussed in more detail next:

Describing People and Social Networks: Friend-of-a-Friend[4] (FOAF) is a vocabulary for describing people, including names, contact information,

[2]For instance, named entity recognition methods typically have 85-90% accuracy on news but only 30-50% on tweets [75,116].

[3]A recently study of 1.1 million tweets has found that 26% of English tweets contain a URL, 16.6% – a hashtag, and 54.8% contain a user name mention [26].

[4]http://xmlns.com/foaf/0.1/

Table 1
Ontologies and what they model

Ontology	People	Online posts	Social networks	Micro blogs	User Interests	Tags	Geo-location	User Behaviour
FOAF	Yes		knows		Partial			
SIOC(T)	Yes	Yes		Partial	Yes			
MOAT						Yes		
Bottari	Yes	Yes	Yes	Yes		Yes	Yes	
DLPO	Yes	Yes	Yes	Yes	Yes	Yes		
SWUM	Yes				Yes		Yes	Yes
UBO	Yes		Yes		Yes			Yes

and a generic `knows` relation. FOAF also supports limited modelling of interests by modelling them as pages on the topics of interest. As acknowledged in the FOAF documentation itself, such an ontological model of interests is somewhat limited.

Modelling Social Media Sites: The Semantically Interlinked Online Communities[5] (SIOC) ontology models social community sites (e.g. blogs, wikis, online forums). Key concepts are forums, sites, posts, user accounts, user groups, and tags. SIOC supports modelling of user interests through the `sioc:topic` property, which has a URI as a value (posts and user groups also have topics).

Modelling microblogs: SIOC has recent extensions (SIOCT), modelling microblogs through the new concept of `MicroblogPost`, a `sioc:follows` property (representing follower/followee relationships on Twitter), and a `sioc:addressed_to` property for posts that mention a specific user name. *Bottari* [28] is an ontology, which has been developed specifically to model relationships in Twitter, especially linking tweets, locations, and user sentiment (positive, negative, neutral), as extensions to the SOIC (Socially-Interlinked Online Communities) ontology. A new TwitterUser class is introduced, coupled with separate `follower` and `following` properties, similar to those in SIOCT. The Tweet class is a type of sioc:Post and, unlike SIOCT, Bottari also distinguishes retweets and replies. Locations (points-of-interest) are represented using the W3C Geo vocabulary[6], which enables location-based reasoning.

Interlinking Social Media, Social Networks, and Online Sharing Practices: DLPO (The LivePost Ontology) provides a comprehensive model of social media posts, going beyond Twitter [124]. It is strongly grounded in fundamental ontologies, such as FOAF, SOIC, and the Simple Knowledge Organisation System (SKOS)[7]. DLPO models personal and social knowledge discovered from social media, as well as linking posts across personal social networks. The ontology captures six main types of knowledge: online posts, different kinds of posts (e.g. retweets), microposts, online presence, physical presence, and online sharing practices (e.g. liking, favouriting). However, while topics, entities, events, and time are well covered, user behaviour roles and individual traits are not addressed as comprehensively as in the SWUM ontology [108] discussed below.

Modelling Tag Semantics: The MOAT (Meaning-Of-A-Tag) ontology [101] allows users to define the semantic meaning of a tag through Linking Open Data and ultimately, to create manually semantic annotations of social media. The ontology defines two kinds of tags: global (across all content) and local (particular tag on a given resource). MOAT can be combined with SIOCT to tag microblog posts [100]. The DLPO ontology, introduce above, also models topics and tags associated with online posts (including microblogs).

User modelling ontologies are key to the representation, aggregation, and sharing of information about users and their social media interactions. The General User Modelling Ontology (GUMO) [59], for instance, aims to cover a wide range of user-related information, such as demographics, contact information, personality, etc. However, it

[5] http://sioc-project.org/
[6] http://www.w3.org/2003/01/geo/
[7] http://www.w3.org/2004/02/skos/. Developed to model thesauri, term lists and controlled vocabularies.

falls short of representing user interests, which makes it unsuitable for social media.

Based on an analysis of 17 social web applications, Plumbaum *et al* [108] have derived a number of user model dimensions required for a social web user modelling ontology. Their taxonomy of dimensions includes demographics, interests and preferences, needs and goals, mental and physical state, knowledge and background, user behaviour, context, and individual traits (e.g. cognitive style, personality). Based on these, they have created the SWUM (Social Web User Model) ontology. A key shortcoming of SWUM, however, is its lack of grounding in other ontologies. For instance, user location attributes, such as Country and City, are coded as strings, which severely limits their usefulness for reasoning (e.g. it is hard to find all users based in South West England, based on their cities). A more general approach would have been to define these through URIs, grounded in commonly used Linked Data resources, such as DBpedia and Freebase.

Lastly, the User Behaviour Ontology [6] models user interactions in online communities. It has been used to model user behaviour in online forums [6] and also Twitter discussions [118]. It has classes that model the impact of posts (replies, comments, etc), user behaviour, user roles (e.g. popular initiator, supporter, ignored), temporal context (time frame), and other interaction information. Addressing the temporal dimension of social media is of particular important, especially when modelling changes over time (e.g. in user interests or opinions).

To summarise, there are a number of specialised ontologies, aimed at representing and reasoning with automatically derived semantic information from social media. However, given that they address different phenomena, many applications adopt or extend more than one, in order to meet their requirements.

4. Semantic Annotation of Social Media

The process of tying semantic models and natural language together is referred to as *semantic annotation*. It may be characterised as the dynamic creation of interrelationships between *ontologies* and unstructured and semi-structured documents in a bidirectional manner [69]. From a technological perspective, semantic annotation is about annotating in texts all mentions of concepts from the ontology (i.e., classes, instances, properties, and relations), through metadata referring to their URIs in the ontology. Approaches which enhance the ontology with new instances derived from texts are typically referred to as *ontology population*. For an in-depth introduction to ontology-based semantic annotation from textual documents see [20].

The automatic semantic annotation of user-generated content enables semantic-based search, browsing, filtering, recommendation, and visual analytics (see Section 6), as well the building of semantic models of the user, their social network, and online behaviour (see Section 5). It is relevant in many application contexts, e.g., knowledge management, competitor intelligence, customer relation management, eBusiness, eScience, eHealth, and eGovernment.

Semantic annotation can be performed manually, automatically, or semi-automatically, i.e., first an automatic system creates some annotations and these are then post-edited and corrected by human annotators.

In the context of social media, the Semantic Microblogging (SMOB) framework has been proposed [100], in order to allow users to add manually machine-readable semantics to messages. SMOB supports also interlinking with the LOD cloud, through hashtags. Hepp [60] proposes a different manual semantic annotation syntax for tweet messages, which is then mapped to RDF statements. The syntax supports relationships between tags (including sameAs), properties from ontologies such as FOAF, and multiple RDF statements in the same tweet.

However, while such manual semantic annotation efforts are valuable, automatic semantic annotation methods are required, in order to make sense of the millions of messages posted daily on Facebook, Twitter, LinkedIn, etc. Consequently, in this section we focus primarily on automatic approaches.

Information Extraction (IE), a form of natural language analysis, is becoming a central technology for bridging the gap between unstructured text and formal knowledge expressed in ontologies. Ontology-Based IE (OBIE) is IE which is adapted specifically for the semantic annotation task [73]. One of the important differences between traditional IE and OBIE is in the use of a formal ontology as one of the system's inputs and as the target output. Some researchers (e.g., [83]) call ontology-based any system which specifies its outputs with respect to an ontology, however, in our view, if a system only has a mapping between the IE outputs

and the ontology, this is not sufficient and therefore, such systems should be referred as *ontology-oriented*.

Another distinguishing characteristic of the ontology-based IE process is that it not only finds the (most specific) class of the extracted entity, but also identifies it, by linking it to its semantic description in the target knowledge base, typically via a URI. This allows entities to be traced across documents and their descriptions to be enriched during the IE process. In practical terms, this requires automatic recognition of named entities, terms, and relations and also co-reference resolution both within and across documents. These more complex algorithms are typically preceded by some shallow linguistic pre-processing (tokenisation, Part-Of-Speech (POS) tagging, etc.)

Linking Open Data resources, especially DBpedia, YAGO and Freebase, have become key sources of ontological knowledge for semantic annotation, as well as being used as target entity knowledge bases for disambiguation. These offer: (i) cross-referenced domain-independent hierarchies with thousands of classes and relations and millions of instances; (ii) an inter-linked and complementary set of resources with synonymous lexicalisations; (iii) grounding of their concepts and instances in Wikipedia entries and other external data. The rich class hierarchies are used for fine-grained classification of named entities, while the knowledge about millions of instances and their links to Wikipedia entries are used as features in the OBIE algorithms.

The rest of this section focuses specifically on methods for semantic annotation of social media streams.

4.1. Keyphrase Extraction

Automatically selected keyphrases are useful in representing the topic of a document or collection of documents, and less effective in delivering arguments or full statements contained therein. Keyphrase extraction can therefore be considered as a form of shallow knowledge extraction, giving a topical overview. Keywords can also be used in the context of semantic annotation and retrieval, as a means of dimensionality reduction and allowing systems to deal with smaller sets of important terms rather than whole documents.

Some keyword extraction approaches exploit term co-occurrence; forming a graph of terms with edges derived from the distance between occurrences of a pair of terms and assigning weights to vertices [91]. This class of keyword extraction was found to perform favourably on Twitter data compared to methods which relied on text models [142].

These graph-based approaches to extracting keywords from Twitter perhaps perform well because the domain contains a great deal of redundancy [146]. For example, in the context of trending topics on Twitter (frequently denoted by hashtags), [127] extracted keyphrases by exploiting textual redundancy and selecting common sequences of words. While redundancy in Twitter and other social media is somewhat beneficial when producing keyword summaries, another less helpful trait is the sheer variety of topics discussed. In cases where documents discuss more than one topic, it can be more difficult to extract a coherent and faithful set of keywords from it.

Personal Twitter timelines, when treated as single documents, present this problem. Users are generally capable of posting on multiple topics. While [142] use TextRank on the whole of a user's stream, they do not attempt to model or address topic variation, unlike [143], who incorporated topic modelling into their approach. Theirs is not the only application of Topic Modelling to Twitter data, as it is similar to [114]. However in the latter work topics are discovered but never summarised.

In the context of social tagging and bookmarking services such as Flickr, Delicious, and Bibsonomy, researchers have studied the automatic tagging of new documents with folksonomy tags. One of the early approaches is the AutoTag system [94], which assigns tags to blog posts. First, it finds similar pre-indexed blog posts using standard information retrieval methods, using the new blog post as the query. Then it composes a ranked list of tags, derived from the top most relevant posts, boosted with information about tags used previously by the given blogger.

More recent approaches use keyphrase extraction from blog content, in order to suggest new tags. For instance, [111] generate candidate keyphrases from n-grams, based on their POS tags, then filter these using a supervised, logistic regression classifier. The keyphrase-based method can be combined with information from the folksonomy [131], in order to generate tag signatures (i.e. associate each tag in the folksonomy with weighted, semantically related terms). These are then compared and ranked against the new blog post, in order to suggest the most relevant set of tags.

Table 2
Ontology-Based Semantic Annotation: Selected Research Tools

	Ontology/ LOD resource used	Annotations produced	Disamb. performed	Target domain	Corpora Used	Evaluated On
DBpedia Spotlight [85]	DBpedia, Freebase	Over 30 classes	Yes	Open domain	Wikipedia	News
LINDEN [128]	YAGO	YAGO classes	Yes	Open domain	Wikipedia	TAC-KBP 2009
Ritter [116]	Freebase	10 classes	No	Open domain	Tweets	Tweets
Ireson [64]	GeoPlanet	Locations	Yes	Photos	Flickr	Flickr
Laniado&Mika [72]	Freebase	Freebase	Yes	Open domain	Tweets	Tweets
Meij [84]	Wikipedia	Wikipedia	Yes	Open domain	Wikipedia	Tweets
Gruhl [54]	MusicBrainz	Songs and albums	Yes	Music domain	MySpace	MySpace posts
Rowe [119]	DBpedia	Conference-related	Yes	Conferences	Tweets	200 tweets
Choudhury [34]	Wikipedia	Cricket players, games	No	Sport events	Wikipedia	Cricket tweets

4.2. Ontology-Based Entity Recognition in Social Media

Ontology-based entity recognition is often broken down into two main phases: entity annotation (or candidate selection) and entity linking (also called reference disambiguation or entity resolution). Ontology-based entity annotation is concerned with identifying all mentions in the text of classes and instances from the ontology (e.g. DBpedia). The entity linking step then uses contextual information from the text, as well as knowledge from the ontology to choose the correct URI. However, it must be noted that not all methods carry out both steps, i.e. some only identify mentions of entities in the text and their class [73].

Table 2 provides an overview of the various approaches discussed in more detail next.

4.2.1. Wikipedia-based Approaches

Most recent work on entity recognition and linking has used Wikipedia as a large, freely available human-annotated training corpus. The target knowledge bases are typically DBpedia [85] or YAGO [128], due to being derived from Wikipedia and thus offering a straightforward mapping between an entity URI and its corresponding Wikipedia page. These more recent, ontology-based approaches have their roots in methods that enrich documents with links to Wikipedia articles (e.g. [93]).

Ontology-based entity disambiguation methods typically collect a dictionary of labels for each entity URI, using the Wikipedia entity pages, redirects (used for synonyms and abbreviations), disambiguation pages (for multiple entities with the same name), and anchor text used when linking to a Wikipedia page. This dictionary is used for identifying all candidate entity URIs for a given text mention. Next is the disambiguation stage, where all candidate URIs are ranked and a confidence score is assigned. If there is no matching entity in the target knowledge base, a NIL value is returned. Text mentions can be disambiguated either independently of each other, or jointly across the entire document (e.g. [93]).

Typically methods use Wikipedia corpus statistics coupled with techniques (e.g. TF/IDF) which match the context of the ambiguous mention in the text against the Wikipedia pages for each candidate entity (e.g. [85]). Michelson *et al* [90] demonstrate how such an approach can be used to derive from a user's tweets, her/his topic profile, which is based on Wikipedia categories. The accuracy of these algorithms has so far been evaluated primarily on Wikipedia articles and news datasets, which are in nature very different from the shorter messages in social media streams.

One widely used Wikipedia-based semantic annotation system is *DBpedia Spotlight* [85]. It is a freely available and customisable web-based system, which annotates text documents with DBpedia URIs. It targets the DBpedia ontology, which has more than 30 top level classes and 272 classes overall. It is possible to restrict which classes (and their sub-classes) are used for named entity recognition, either by listing them explicitly or through a SPARQL query. The algorithm first selects entity candidates through lookup against a Wikipedia-derived dictionary of URI lexicalisations, followed by a URI ranking stage using a vector space model. Each DBpedia resource is associated with a document, constructed from all paragraphs mentioning that concept in Wikipedia. The method has been shown to out-perform OpenCalais and Zemanta (see Section 4.2.3) on a small gold-standard of newspaper articles [85].

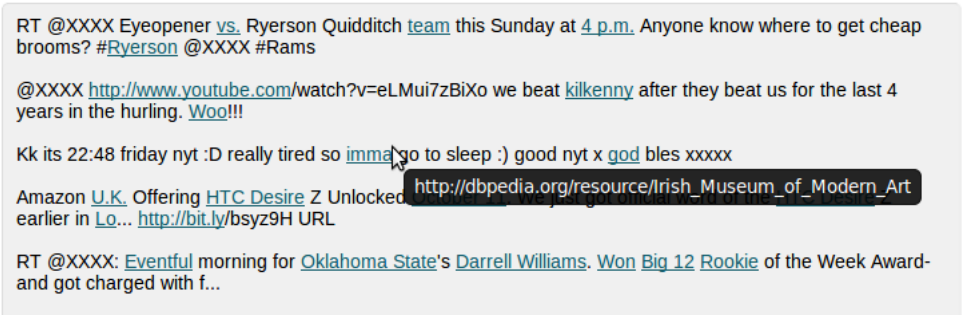

Fig. 1. DBpedia Spotlight results on tweets

Figure 1 shows several tweets annotated with DBpedia Spotlight. The results clearly demonstrate the need for tweet spelling normalisation, as well as the difficulties Spotlight has with recognising URLs. As exemplified here, by default the algorithm is designed to maximise recall (i.e. annotate as many entities as possible, using the millions of instances from DBpedia). Given the short, noisy nature of tweets, this may lead to low accuracy results. Further formal evaluation on a shared, large dataset of short social media messages is required, in order to establish the best values for the various DBpedia Spotlight parameters (e.g. confidence, support).

The LINDEN [128] framework makes use of the richer semantic information in YAGO (semantic similarity), in addition to Wikipedia-based information (using link structure for semantic associativity). The method is heavily dependent on the Wikipedia-Miner[8] toolkit [93], which is used to analyse the context of the ambiguous entity mention and detect the Wikipedia concepts that appear there. Evaluation on the TAC-KBP2009 dataset showed LINDEN outperforming the highest ranked Wikipedia-only systems, which participated in the original TAC evaluation. Unfortunately, LINDEN has not been compared directly to DBpedia Spotlight on a shared evaluation dataset.

4.2.2. Social Media Oriented Approaches

Named entity recognition methods, which are typically trained on longer, more regular texts (e.g. news articles), have been shown to perform poorly on shorter and noisier social media content [116]. However, while each post in isolation provides insufficient linguistic context, additional information can be derived from the user profiles, social networks, and interlinked posts (e.g. replies to a tweet message). This section discusses what we call *social media oriented* semantic annotation approaches, which integrate both linguistic and social media-specific features.

Ritter *et al* [116] address the problem of named entity classification (but not disambiguation) by using Freebase as the source of large number of known entities. The straightforward entity lookup and type assignment baseline, without considering context, achieves only 38% f-score (35% of entities are ambiguous and have more than one type, whereas 30% of entities in the tweets do not appear in Freebase). NE classification performance improves to 66% through the use of labelled topic models, which take into account the context of occurrence and the distribution over Freebase types for each entity string (e.g. Amazon can be either a company or a location).

Ireson *et al* [64] study the problem of location disambiguation (toponym resolution) of name tags in Flickr. The approach is based on the Yahoo! GeoPlanet semantic database, which provides a URI for each location instance, as well as a taxonomy of related locations (e.g. neighbouring locations). The tag disambiguation approach makes use of all other tags assigned to the photo, the user context (all tags assigned by this user to all their photos), and the extended user context, which takes into account the tags of the user contacts. The use of this wider, social network-based context was shown to improve significantly the overall disambiguation accuracy.

Another source of additional, implicit semantics are hashtags in Twitter messages, which have evolved as means for users to follow conversations on a given topic. Laniado and Mika [72] investigate hashtag semantics in 369 million messages, using four metrics: frequency of use, specificity (use of the hashtag vs use of the word itself), consistency of usage, and stability over time. These measures are then used to determine which hashtags can be used as identifiers and

[8] http://wikipedia-miner.cms.waikato.ac.nz/

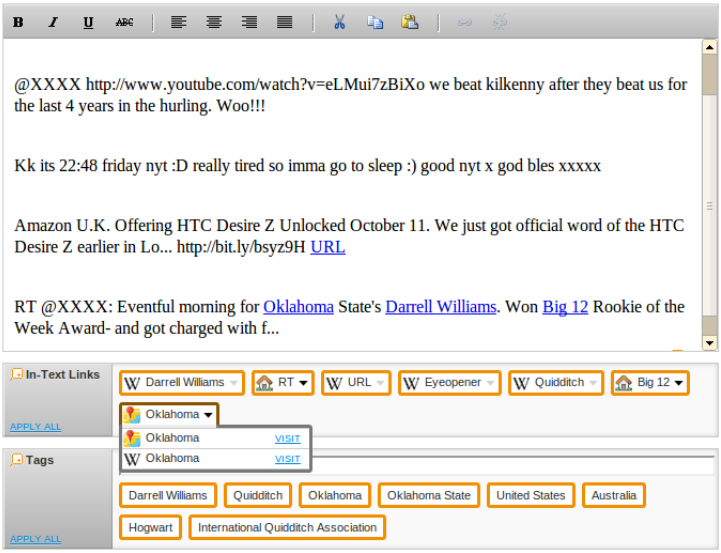

Fig. 2. Zemanta's online tagging interface

linked to Freebase URIs (most of them are named entities). Hashtags have also been used as an additional source of semantic information about tweets, by adding textual hashtag definitions from crowdsourced online glossaries [87]. Mendes *et al* [87] also carry out semantic annotation through a simple entity lookup against DBpedia entities and categories without further disambiguation. User-related attributes and social connections are coded in FOAF, whereas semantic annotations are coded through the MOAT ontology (see Section 3).

Wikipedia-based entity linking approaches (see Section 4.2.1) benefit significantly from the larger linguistic context of news articles and web pages. Evaluation of DBpedia Spotlight [85] and the Milne and Witten method [93] on a tweet dataset has shown significantly poorer performance [84]. Meij *et al* [84] propose a Twitter-specific approach for linking such short, noisy messages to Wikipedia articles. The first step uses n-grams to generate a list of candidate Wikipedia concepts, then supervised learning is used to classify each concept as relevant or not (given the tweet and the user who wrote it). The method uses features derived from the n-grams (e.g. number of Wikipedia articles containing this n-gram), Wikipedia article features (e.g. number of articles linking to the given page), and tweet-specific features (e.g. using hashtag definitions and linked web pages).

Gruhl *et al.* [54] focus in particular on the disambiguation element of semantic annotation and examine the problem of dealing with highly ambiguous cases, as is the case with song and music album titles. Their approach first restricts the part of the MusicBrainz ontology used for producing the candidates (in this case by filtering out all information about music artists not mentioned in the given text). Secondly, they apply shallow language processing, such as POS tagging and NP chunking, and then use this information as input to a support vector machine classifier, which disambiguates on the basis of this information. The approach was tested on a corpus of MySpace posts for three artists. While the ontology is very large (thus generating a lot of ambiguity), the texts are quite focused, which allows the system to achieve good performance. As discussed by the authors themselves, the processing of less focused texts, e.g. Twitter messages or news articles, is likely to prove much more challenging.

4.2.3. Commercial Entity Recognition Services

There are a number of commercial online entity recognition services which annotate documents with entities and assign Linked Data URIs to them. The NERD online tool [117] allows their easy comparison on user-uploaded datasets. It also unifies their results and maps them to the Linking Open Data cloud. Here we focus only on the services used by research methods surveyed here (e.g. [121,2,119]).

Zemanta (http://www.zemanta.com) is an online semantic annotation tool, originally developed for blog and email content to help users insert tags and links through recommendations. Figure 2 shows an example text and the recommended tags, potential in-text

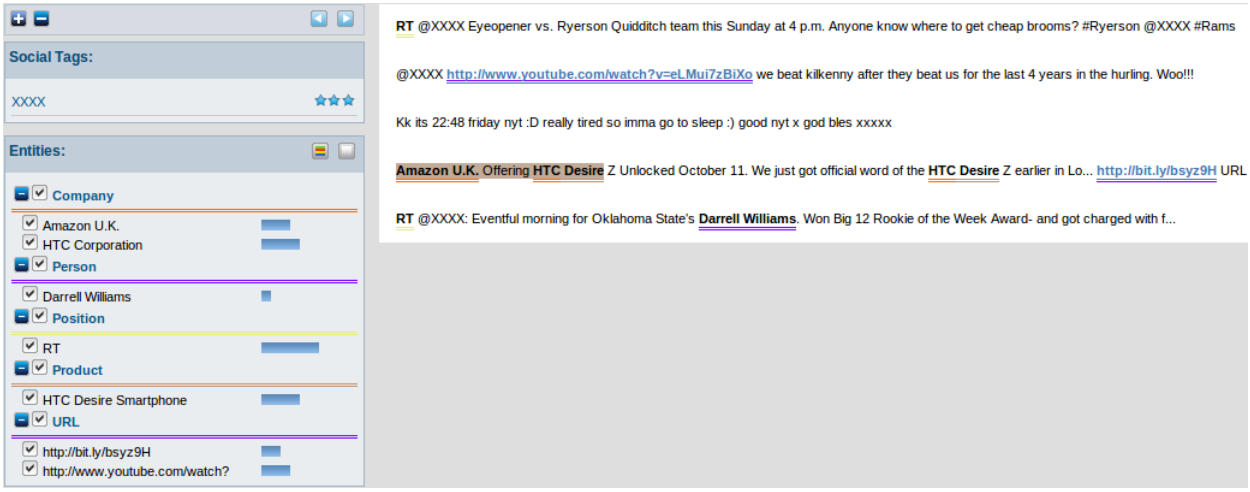

Fig. 3. Calais results on tweets

link targets (e.g., the W3C Wikipedia article and the W3C home page), and other relevant articles. It is then for the user to decide which of the tags should apply and which in-text link targets they wish to add. In this example, in-text links have been added for the terms highlighted in orange, all pointing to the Wikipedia articles on the respective topics.

Open Calais is another commercial web service for semantic annotation, which has been used by some researchers on social media. For instance, Abel *et al* [2] harness OpenCalais to recognise named entities in news-related tweets[9]. The target entities are mostly locations, companies, people, addresses, contact numbers, products, movies, etc. The events and facts extracted are those involving the above entities, e.g., acquisition, alliance, company competitor. Figure 3 shows an example text annotated with some entities.

The entity annotations include URIs, which allow access via HTTP to obtain further information on that entity via Linked Data. Currently OpenCalais links to eight Linked Data sets, including its own knowledge base, DBpedia, Wikipedia, IMDB and Shopping.com. These broadly correspond to the entity types covered by the ontology.

The main limitation of Calais comes from its proprietary nature, i.e., users send documents to be annotated by the web service and receive results back, but they do not have the means to give Calais a different ontology to annotate with or to customise the way in which the entity extraction works.

[9]Unfortunately they do not evaluate the named entity recognition accuracy of OpenCalais on their dataset.

4.3. Event Detection

Much as trending topics can be used to monitor global opinions and reactions, social media streams can be used as a discussion backchannel to real world events [41], and even to discover and report upon such events, almost as soon as they occur. While it may at first appear that trending topics alone are sufficient for this task, there are a few reasons why they are unsatisfactory:

- *Generality*: trending topics may discuss events, but may also refer to celebrities, products or online memes.
- *Scale*: only the topics with which a huge margin of Twitter users engage can appear as trending topics.
- *Censorship*: it is believed by many that the trending topics displayed by the official Twitter service are censored for political and language content.
- *Algorithm*: the method used to select trending topics is not published anywhere and is generally not understood.

Automatic event detection therefore presents an interesting task for social media streams. While it is possible to have access to an enormous quantity of tweets, enough to reveal global trends and events, the problem of developing and evaluating scalable event detection algorithms which can handle such magnitudes of streaming text remains.

The majority of approaches to event detection do not utilise ontologies or other sources of semantic information. One class of methods use clustering on tweets [105,14,15] or blog posts [123]. Another class of event detection methods take inspiration from signal processing, analysing tweets as sensor data. [122] used such an approach to detect earthquakes in Japan on the basis of Tweets with geolocation information attached to them. Similarly, individual words have been treated as wavelet signals and analysed as such in order to discover temporally significant clusters of terms [141].

Once an event is detected in social media streams, the next problem is how to generate useful thematic/topical descriptors for this event. Point-wise mutual information has been coupled with user geolocation and temporal information, in order to derive n-gram event descriptors from tweets [96]. By making the algorithm sensitive to the originating location, it is possible to see what people from a given location are saying about an event (e.g. those in the US), as well as how this differs from tweets elsewhere (e.g. those from India).

Collections of events in a larger sequence could be referred to as sagas; they may be perfectly legitimate events in their own right, or their individual constituents might similarly be coherent on their own. Citing the example of an academic conference, [119] point out that tweets may refer to the conference as a whole, or to specific sub-events such as presentations at a specific time and place. Using semantic information about the conference event and its sub-events from the Web of Data, tweets are aligned to these sub-events automatically, using machine learning. The method includes a concept enrichment phase, which uses Zemanta to annotate each tweet with DBpedia concepts. Tweets are described semantically using the SIOC and Online Presence semantic ontologies (see Section 3).

Another semantic, entity-based approach to sub-event detection has been proposed by [34], who use manually created background knowledge about the event (e.g. team and player names for cricket games), coupled with domain-specific knowledge from Wikipedia (e.g. cricket-related sub-events like getting out). In addition to annotating the tweets with this semantic information, the method utilises tweet volume (similarly to [78]) and re-tweet frequency as sub-event indicators. The limitation of this approach, however, comes from the need for manual intervention, which is not always feasible outside of limited application domains.

4.4. Sentiment Detection and Opinion Mining

The existence and popularity of websites dedicated to reviews and feedback on products and services is something of a homage to the human urge to post what they feel and think online. When the most common type of message on Twitter is about 'me now' [95], it is to be expected that users talk often about their own moods and opinions. Bollen *et al* [19] argue that users express both their own mood in tweets about themselves and more generally in messages about other subjects. Another study [66] estimates that 19% of microblog messages mention a brand and from those that do, around 20% contain brand sentiment.

The potential value of these thoughts and opinions is enormous. For instance, mass analysis could provide a clear picture of overall mood, exploring reactions to ongoing public events [19] or feedback to a particular individual, government, product or service [71]. The resulting information could be used to improve services, shape public policy or make a profit on the stock market.

The user activities on social networking sites are often triggered by specific events and related entities (e.g. sports events, celebrations, crises, news articles, persons, locations) and topics (e.g. global warming, financial crisis, swine flu). In order to include this information, semantically- and social network-aware approaches are needed.

There are many challenges inherent in applying typical opinion mining and sentiment analysis techniques to social media [81]. Microposts are, arguably, the most challenging text type for opinion mining, since they do not contain much contextual information and assume much implicit knowledge. Ambiguity is a particular problem since we cannot easily make use of coreference information: unlike in blog posts and comments, tweets do not typically follow a conversation thread, and appear much more in isolation from other tweets. They also exhibit much more language variation, tend to be less grammatical than longer posts, contain unorthodox capitalisation, and make frequent use of emoticons, abbreviations and hashtags, which can form an important part of the meaning. Typically, they also contain extensive use of irony and sarcasm, which are particularly difficult for a machine to detect. On the other hand, their terseness can also be beneficial in focusing the topics more explicitly: it is very rare for a single tweet to be related to more than one topic, which can thus aid disambiguation by emphasising situational relatedness.

[99] present a wide-ranging and detailed review of traditional automatic sentiment detection techniques, including many sub-components, which we shall not repeat here. In general, sentiment detection techniques can be roughly divided into lexicon-based methods (e.g. [125,135]) and machine-learning methods, e.g. [18]. Lexicon-based methods rely on a sentiment lexicon, a collection of known and pre-compiled sentiment terms. Machine learning approaches make use of shallow syntactic and/or linguistic features [98,52], and hybrid approaches are very common, with sentiment lexicons playing a key role in the majority of methods, e.g. [38].

The majority of sentiment and opinion mining methods tested on social media utilise no or very little semantics. For instance, [98,52] classify tweets as having positive, negative, or neutral sentiment, based on n-grams and part-of-speech information, whereas [38] use a sentiment lexicon to initially annotate positive and negative sentiment in tweets related to political events.

The use of such shallow linguistic information leads to a data sparsity problem. Saif *et al* [121] demonstrate that by using semantic concepts, instead of words such as iPhone, polarity classification accuracy is improved. The approach uses AlchemyAPI for semantic annotation of 30 entity classes, the most frequent ones being Person, Company, City, Country, and Organisation. The method is evaluated on the Stanford Twitter Sentiment Dataset[10] and shown to outperform semantics-free, state-of-the-art methods, including [52].

Semantic annotation has also been used for the more challenging opinion mining task. In particular, [79] identify people, political parties, and opinionated statements in tweets using a rule-based entity recogniser, coupled with an affect lexicon derived from WordNet. Subsequent semantic analysis uses patterns to generate triples representing opinion holders and voter intentions. Negation is dealt with by capturing simple patterns such as 'isn't helpful' or 'not exciting' and using them to negate the extracted sentiment judgements.

4.5. Cross-Media Linking

The short nature of Twitter and Facebook messages, coupled with their frequent grounding in real world events, means that often short posts cannot be understood without reference to external context. While some posts already contain URLs, the majority do not. Therefore automatic methods for cross-media linking and enrichment are required.

Abel *et al* [2] link tweets to current news stories in order to improve the accuracy of semantic annotation of tweets. Several linkage strategies are explored: utilising URLs contained in the tweet, TF-IDF similarity between tweet and news article, hashtags, and entity-based similarity (semantic entities and topics are recognised by OpenCalais), with the entity-based one being the best one for tweets without URLs. The approach bears similarities with the keyphrase-based linking strategy for aligning news video segments with online news pages [42]. [62] go one step further by aggregating social media content on climate change from Twitter, YouTube, and Facebook with online news, although details of the cross-media linking algorithm are not supplied in the paper.

An in-depth study comparing Twitter and New York Times news [148] has identified three types of topics: event-oriented, entity-oriented, and long-standing topics. Topics are also classified into categories, based on their subject area. Nine of the categories are those used by NYT (e.g. arts, world, business) plus two Twitter-specific ones (Family&Life and Twitter). Family&Life is the predominant category on Twitter (called 'me now' by [95]), both in terms of number of tweets and number of users. Automatic topic-based comparison showed that tweets abound with entity-oriented topics, which are much less covered by traditional news media.

Going beyond news and tweets, future research on cross-media linking is required. For instance, some users push their tweets into their Facebook profiles, where they attract comments, separate from any tweet replies and retweets. Similarly, comments within a blog page could be aggregated with tweets discussing it, in order to get a more complete overall view.

4.6. Discussion

Even though some inroads have been made already, current methods for semantic annotation of social media streams have many limitations. Firstly, most methods address the more shallow problems of keyword and topic extraction, while ontology-based entity and event recognition do not reach the significantly higher precision and recall results obtained on longer text documents. One way to improve the currently poor automatic performance is through crowdsourcing. The ZenCrowd system [37], for instance, combines algo-

[10] http://twittersentiment.appspot.com/

rithms for large-scale entity linking with human input through micro-tasks on Amazon Mechanical Turk. In this way, textual mentions that can be linked automatically and with high confidence to instances in the LOD cloud, are not shown to the human annotators. The latter are only consulted on hard to solve cases, which not only significantly improves the quality of the results, but also limits the amount of manual intervention required. We return to crowdsourcing in more detail in Section 7.

Another way to improve semantic annotation of social media is to make better use of the vast knowledge available on the Web of Data. Currently this is limited mostly to Wikipedia and resources derived from it (e.g. DBpedia and YAGO). One of the challenges here is ambiguity. For instance, song and album titles in MusicBrainz are highly ambiguous and include common words (e.g. Yesterday), as well as stop words (The, If) [54]. Consequently, an automatic domain categorisation step might be required, in order to ensure that domain-specific LOD resources, such as MusicBrainz, are used to annotate only social media content from the corresponding domain. The other major challenges are robustness and scalability. Firstly, the semantic annotation algorithms need to be robust in the face of noisy knowledge in the LOD resources, as well as being robust with respect to dealing with the noisy, syntactically irregular language of social media. Secondly, given the size of the Web of Data, designing ontology-based algorithms which can load and query efficiently these large knowledge bases, while maintaining high computational throughput is far from trivial.

The last obstacle to making better use of Web of Data resources, lies in the fairly limited lexical information available. With the exception of resources grounded in Wikipedia, lexical information in the rest is mostly limited to RDF labels. This in turn limits their usefulness as a knowledge source for ontology-based information extraction and semantic annotation. One recent strand of work has focused on utilising the Wiktionary [89] collaboratively built, multilingual lexical resources. It is particularly relevant to analysing user-generated content, since it contains many neulogisms and is updated continuously by its contributor community. For English and German, in particular, there is also related ongoing work on creating UBY [55] – a unified, large-scale, lexico-semantic resource, grounded in Wikipedia and Wordnet, and thus, indirectly, to other LOD resources as well. Another relevant strand is work on linguistically grounded ontologies [22], which has proposed a more expressive model for associating linguistic information to ontology elements. While these are steps in the right direction, nevertheless further work is still required, especially with respect to building multilingual semantic annotation systems.

In addition, it is axiomatic that semantic annotation methods are only as good as their training and evaluation data. Algorithm training on social media gold standard datasets is currently very limited. For example, there are currently fewer than 10,000 tweets annotated with named entity types and events. Bigger, shared evaluation corpora from different social media genres are therefore badly needed. Creating these through traditional manual text annotation methodologies is unaffordable, if a significant mass is to be reached. Research on crowdsourcing evaluation gold standards has been limited, primarily with focus on using Amazon Mechanical Turk to acquire small datasets (e.g. tweets with named entity types) [47]. We will revisit this challenge again in Section 7.

In the area of sentiment analysis, researchers have investigated the problems of sentiment polarity detection, subjectivity classification, prediction through social media and user mood profiling, however, most methods use no or very little semantics. Moreover, evaluation of opinion mining is particularly difficult for a number of methodological reasons (in addition to the lack of shared evaluation resources discussed above). First, opinions are often subjective, and it is not always clear what was intended by the author. For example, a person cannot necessarily tell if a comment such as "I love Baroness Warsi", in the absence of further context, expresses a genuine positive sentiment or is being used sarcastically. Inter-annotator agreement performed on manually annotated data therefore tends to be low, which affects the reliability of any gold standard data produced.

Lastly, social media streams impose a number of further outstanding challenges on opinion and sentiment mining methods:

- *Relevance*: In social media, discussions and comment threads can rapidly diverge into unrelated topics, as opposed to product reviews which rarely stray from the topic at hand.
- *Target identification*: There is often a mismatch between the topic of the social media post, which is not necessarily the object of the sentiment held therein. For example, the day after Whitney Houston's death, TwitterSentiment and similar sites all showed an overwhelming majority

of tweets about Whitney Houston to be negative; however, almost all these tweets were negative only in that people were sad about her death, and not because they disliked her.

- *Volatility over time*: More specifically, opinions can change radically over time, from positive to negative and vice versa. To address this problem, the different types of possible opinions can be associated as ontological properties with the classes describing entities, facts and events, discovered through semantic annotation techniques, similar to those in [80] which aimed at managing the evolution of entities over time. The extracted opinions and sentiments can be time-stamped and stored in a knowledge base, which is enriched continuously, as new content and opinions come in. A particularly challenging question is how to detect emerging new opinions, rather than adding the new information to an existing opinion for the given entity. Contradictions and changes also need to be captured and used to track trends over time, in particular through opinion aggregation.
- *Opinion aggregation*: Another challenge is the type of aggregation that can be applied to opinions. In entity-based semantic annotation, this can be applied to the extracted information in a straightforward way: data can be merged if there are no inconsistencies, e.g. on the properties of an entity. Opinions behave differently here, however: multiple opinions can be attached to an entity and need to be modelled separately, for which we advocate populating a knowledge base. An important question is whether one should just store the mean of opinions detected within a specific interval of time (as current opinion visualisation methods do), or if more detailed approaches are preferable, such as modelling the sources and strength of conflicting opinions and how they change over time. A second important question in this context involves finding clusterings of the opinions expressed in social media, according to influential groups, demographics and geographical and social cliques. Consequently, the social, graph-based nature of the interactions requires new methods for opinion aggregation.

However, even though state-of-the-art methods have a large scope for improvement, semantic annotation results are already being used by methods that automatically derive models of users and social networks, from the information implicit in social media streams. This is where we turn to next.

5. Semantic-Based User Modelling

A *User Model* (UM) is a knowledge resource containing *explicit* semantic information about various aspects of the user, which the system has *a priori* (e.g. by importing a Facebook profile) or has inferred from user behaviour, user-generated content, social networks or other sources. Some important characteristics of user models are:

- UM is a *distinct knowledge resource* within the overall system;
- semantic information is represented *explicitly*. Implicit information disclosed in social media is used to derive this explicit knowledge.
- *abstraction*, i.e., representation of types of users, roles and groups, as well as of individual users.
- *multi-purpose* – the semantically encoded user model can be used in different ways, e.g. personalised content recommendation, filtering.
- *reasoning* – the representation should allow for reasoning *about* the knowledge, as well as reasoning *with* it.
- *interconnected* – a user model is more than a collection of attributes. Usually, there are also complex relations between them, as well as relations to other types of knowledge (e.g. posts made by the user).

Ontology-based user models have been used extensively on content other than social media streams, especially in the context of Personal Information Management (PIM). PIM work originated in research on the social semantic desktop [36], where information from the user's computer (e.g. email, documents) is used to derive models of the user. For a detailed overview of user modelling for the semantic web see [8].

In this paper, we focus on the extension of this work towards social media streams, as well as mention sensor-based information where relevant (e.g. GPS coordinates in tweets). As discussed in Section 2, the social and user-generated nature of these streams make it possible to derive rich semantic user models. More specifically, we examine the application of semantic annotation for user model construction. Consequently, we consider outside the scope of this paper research which is focused purely on social network analysis (e.g. [92]) and/or uses purely quantitative user and post characteristics (e.g. number of threads/posts, number of replies/re-tweets [118]) and/or post metadata only (e.g. the re-tweet and in-reply-to JSON fields).

5.1. Constructing Social Semantic User Models from Semantic Annotations

Among the various kinds of social media, folksonomies have probably received most attention from researchers studying how semantic models of user interactions and interests can be derived from user-generated content. Many approaches focused on exploring the social and interaction graphs, using techniques from social network analysis (e.g. [92]). In this section, however, we are concerned with methods that discover and exploit the semantics of textual tags instead (including hashtags). This section also includes semantic-based user modelling research on online forums, blogs, and Twitter.

Based on the kinds of semantic information used, methods can be classified as follows:

- Bag of words ([31]);
- Semantically disambiguated entities: mentioned by user (e.g. [2,67]) or from a linked longer Web document (e.g. [2]);
- Topics: Wikipedia categories (e.g. [2,133]), latent topics (e.g. [147]), or tag hierarchies (e.g. [149]). In order to model tag semantics more explicitly, researchers [27] have proposed grounding tags into WordNet and then using WordNet-based semantic similarity measures, to derive the semantic relatedness of folksonomy tags.

This is typically supplemented with more quantitative social network information (e.g. how many connections/followers a user has [6]) and interaction information (e.g. post frequency [118], average number of posts per thread [6]).

The rest of the section discusses in more detail the kinds of user information that have been extracted from semantically annotated social media, and concludes with a discussion of open issues.

5.1.1. Discovering User Demographics

Every Twitter user has a profile which reveals some details of their identity. The profile is semi-structured, including a textual bio field, a full name, the user's location, a profile picture, a time zone and a homepage URL (most of these are optional and often empty). The user's attributes can be related to the content of their posts, for example their physical location can determine to a degree the language they use [33] or the events on which they comment [144].

There have been efforts to discover user demographics information, when it is not available in the fields in their profile. [23] classify users as male or female based on the text of their tweets, their description fields and their names. They report better-than-human accuracy, compared to a set of annotators on Mechanical Turk. [103] present a general framework for user classification which can learn to automatically discover political alignment, ethnicity and fans of a particular business.

Twitter users may share the location from which they are tweeting by posting from a mobile device and allowing it to attach a reading from its GPS receiver to the message, or by setting their own location in a field of their profile. However, [33] found that only around 36% of users actually filled in their location field in their profile with a valid location as specific as their nearest city. Furthermore, when we analysed a dataset of over 30,000 tweets discussing the 2011 London Riots, less than 1% of messages contained any GPS information.

There have been attempts to automatically locate Twitter users. Content-based methods ('you are where you write about') typically gather the textual content produced by the user and infer their location based on features, such as mentions of local place names [48] and use of local dialect. In the work of [44,33], region-specific terms and language that might be relevant to the geolocation of users were discovered automatically. A classification approach is devised in [77] that also incorporates specific mentions of places near to the user. One disadvantage to this method is the fact that someone might be writing about a popular global event which is of no relevance to their actual location. Another is that users might take deliberate steps to hide their true location by alternating the style of their posts or not referencing local landmarks.

In contrast, network-based geolocation methods ('you are where your friends are') aim to use the user's social network to infer their location. To the best of our knowledge, the only existing method of this kind (i.e., relying on the user's social network alone) is the work of [9], who first create a model for the distribution of distance between pairs of friends, before using this to find the most likely location for a given user. The influence of distance on social network ties is demonstrated by the earlier work of [74]. The main disadvantage of their approach is that it assumes that all users globally have the same distribution of friends in terms of distance. Also, they do not account explicitly for the density of people in an area.

We have developed a method for collecting a large dataset of users with known, ground truth locations,

which is based on semantically disambiguating the user defined 'location' field in their Twitter profile by assigning the corresponding DBpedia URI. A thorough analysis of how users use this field is presented in [58]. Other work has relied instead on small amounts of geotagged data (e.g FourSquare checkins) as an extra feature for user location or as the only way of locating users. The work of [77] uses these checkins as part of their classification. Although this often leads to high accuracy results, the approach is limited by the very small amount of geo-tagged data available, mainly due to practical constraints (e.g. battery usage) and privacy concerns.

5.1.2. Deriving User Interests

Abel *et al* [2] propose simple entity-based and topic-based user profiles, built from the user's tweets. The entity-based profile for a given user is modelled as a set of weighted entities, where the weight each entity *e* is computed based either on the number of user tweets that mention *e*, or based on frequency of entity occurrences in the tweets, combined with the related news articles (which are identified in an earlier, linking step). Topic-based profiles are defined in a similar fashion, but represent higher level Wikipedia categories (e.g. sports, politics). Both entities and topics are identified using OpenCalais (see Section 4.2.3). Abel *et al* have also demonstrated that hashtags are not a useful indicator of user interests – a finding which is also supported by [90]. A major limitation of the method is that it depends heavily on the news linking, which the authors have shown applies successfully to only 15% of tweets.

In a subsequent paper [3], Abel *et al* refine their approach to modelling user interests in a topic, to take also into account re-tweets, as well as changes over time (when users become interested in a topic, for how long, and which concepts are relevant to which topic). Evaluation is based on global topics (e.g. the Egyptian revolution). Their findings demonstrate that a time-dependent topic weighting function produces user interest models, which are better for tweet recommendation purposes. They also identify different groups of users, based on the duration of their interest in a given topic: *long-term adopters* who join early for longer vs. *short-term adopters* who join global discussions later and are influenced by public trends.

Kapanipathi *et al* [67] similarly use semantic annotations to derive user interests (entities or concepts from DBpedia), weighted by strength (calculated on the basis of frequency of occurrence). They also demonstrate how interests can be merged based on information from different social media (LinkedIn, Facebook and Twitter). Facebook likes and explicitly stated interests in LinkedIn and Facebook are combined with the implicit interest information from the tweets. The Open Provenance Model[11] is used to keep track of interest provenance.

A similar entity- and topic-based approach to modelling user interests is proposed by Michelson and Macskassy [90] (called Twopics). All capitalised, non-stop words in a tweet are considered as entity candidates and looked up against Wikipedia (page titles and article content). A disambiguation step then identifies the Wikipedia entity which matches best the candidate entity from the tweet, given the tweet content as context. For each disambiguated entity, the sub-tree of Wikipedia categories is obtained. In a subsequent, topic-assignment step, all category sub-trees are analysed to discover the most frequently occurring categories, which are then assigned as user interests in the topic-based profile. The authors also argue that such more generic topics, generated by leveraging the Wikipedia category taxonomy, are more appropriate for clustering and searching for users, than the term-based topic models derived using bag-of-words or LDA methods.

Researchers have also investigated the problem of deriving user interests from tags and other metadata in folksonomies. For instance, [17] build a shallow model of user interests from the user-created folksonomy tags and the words appearing in other user-authored metadata (e.g. title, description). This user profile is then used to recommend items from the folksonomy (e.g. Del.icio.us URLs, Bibsonomy articles).

[147] propose a topic-based probabilistic method for identifying user interests from folksonomy tags (Del.icio.us). The first step is to induce hierarchies of latent topics from a set of tags in an unsupervised manner. This approach, based on Hierarchical Dirichlet Process, models topics as probability distributions over the tag space, rather than clustering the tags themselves [149]. Next, user interest hierarchies are induced via log-likelihood and hierarchy comparison methods. Zavitsanos *et al* however stop short of assigning explicit semantics to the topics through URIs.

In order to ground folksonomy tags semantically, Cattuto *et al* [27] mapped pairs of tags in Del.icio.us to pairs of synsets in Wordnet. Then WordNet-based

[11] http://openprovenance.org

measures for semantic distance are used to derive semantic relations between the mapped tags. The researchers demonstrated that a semantically-sound and computationally affordable metric for semantic similarity between tags is the tag context similarity, which measures the tag's co-occurrence with the 10,000 most popular tags in the folksonomy. In this way, tags which belong to the same semantic concept can be identified. The same approach could, in theory, be applied to hashtags in tweets, although its effectiveness remains to be proven.

Others [133] have used Wikipedia as a multi-domain model, that can be used to model semantically user interests. They also propose a method for consolidation of user profiles across social networking sites. Tags from different sites are filtered based on WordNet synonymy and correlated to Wikipedia pages. Subsequently, Wikipedia categories are used, in order to select representative higher-level topics of interest for the user. The approach is very similar to Twopics [90].

Researchers have also demonstrated a link between the kinds of tags and content created and user behaviour categories (see Section 5.1.3), which has direct implications on how well we can derive user interests and/or recommend content. In the context of capturing tag semantics in folksonomies, previous work [70,132] has shown that different users of the same social tagging system can have different tagging motivation (categorisation vs description), which in turn influences the kinds of tags entered in the system. In particular, for the purposes of discovering emergent tag semantics, it is more beneficial to use as input the more prolific and descriptive tags produced by describer users. Going beyond tags, Naaman *et al* [95] have shown that the two different categories of Twitter users (meformers and informers) produce significantly different kinds of tweet content. For instance, informers post primarily information sharing messages, whereas meformers write mainly about themselves and voice opinions and complaints.

5.1.3. Capturing User Behaviour

As demonstrated above, user behaviour is key to understanding interactions in social media. In this section we focus primarily on approaches which utilise automatically-derived semantics, in order to classify user behaviour.

In the case of online forums, the following user behaviour roles have been identified [29]: *elitist*, *grunt*, *joining conversationalist*, *popular initiator*, *popular participant*, *supporter*, *taciturn*, and *ignored*. For social tagging systems, researchers [132] have classified users according to their tagging motivation, into *categorisers* and *describers*. In Twitter, the most common role distinction is drawn on the basis of tweet content and users are classified into *meformers* (80% of users) and *informers* (20% of users) [95].

In order to assign behaviour roles in online forums automatically, Angeletou *et al* [6] create skeleton rules in SPARQL, that map semantic features of user interaction to a level of behaviour (high, medium, and low). These levels are constructed dynamically from user exchanges and can be altered over time, as the communities evolve. User roles, contexts, and interactions are modelled semantically through the User Behaviour Ontology (see Section 3) and are used ultimately to predict the health of a given online forum.

The problem of characterising Twitter user behaviour, based on the content of their posts has yet to be fully explored. [143] generated keyphrases for users with the aid of topic modelling and a PageRank method. Similarly, [142] use a combination of POS filtering and TextRank to discover tags for users. It should also be noted that while [95] went some way towards categorising user behaviour and tweet intention, their method is not automatic and it remains unclear whether or not similar categories could be assigned by a classifier.

5.2. Discussion

As demonstrated by our survey, a key research challenge for semantic user modelling lies in addressing the diverse, dynamic, temporal nature of user behaviour. An essential part of that is the ability to represent and reason with conflicting personal views, as well as to model change in user behaviour, interests, and knowledge over time. For instance, in the context of blogs, Cheng *et al* [32] have proposed an interest forgetting function for short-term and long-term interest modelling. Angeletou *et al* [6] recently developed time-contextualised models of user behaviour and demonstrated how these could be used to predict changes in user participation in online forums.

With respect to tweets, automatically derived user interests could also be separated into "global" ones (based on the user's tweets on trending topics) versus "user-specific" (topics which are of more personal interest, e.g. work, hobby, friends). Further work is required on distinguishing globally interesting topics (e.g. trending news) from interests specific to the given user (e.g. work-related, hobby, gossip from a friend,

etc.). In other words, we need to go beyond modelling what is interesting to a user, to capture also why it is of interest. In more detail, a given tweet could be interesting to a user for social reasons (e.g. my brother posted them), cultural reasons, topical relevance (e.g. match my hobby), or be part of a larger sequence of tweets, forming a conversation. Current methods (see Section 5.1.2) have largely focused on topically relevant tweets, leaving ample scope for future research.

There is also need for further research into modelling how user interests change over time, going beyond the work of Abel *et al* [3], which focused on global topics derived from tweets linked to political news. One challenge is to establish how well the method generalises to tweets from domains other than news, as well as tweets without URLs[12] The latter are likely to prove particularly challenging, since the method most likely benefits significantly from the additional, longer content of the URL. Moreover, as already discussed, tweets can be interesting for a number of reasons, other than global importance. We hypothesize that the interestingness of a given tweet is likely to decay or change differently over time, depending on the reason(s) that make it interesting to a given user.

What is interesting to a user also ties in with user behaviour roles (see Section 5.1.3). In turn, this requires more sophisticated methods for automatic assignment of user roles, based on the semantics of posts, in addition to the current methods based primarily on quantitative interaction patterns.

Since many users now participate in more than one social network, the issue of merging user modelling information across different sources arises, coupled with the challenge of modelling and making use of provenance information. As discussed in Section 5.1.2 above, there has been some preliminary work [67] on merging implicit interests derived from the user's tweets with explicit interests, given on LinkedIn and Facebook. The method currently gives equal weights to the three social sites used to derive the interests. However, more sophisticated models could be derived, for instance, giving higher weights to professional interests derived from LinkedIn for Twitter users who tweet predominantly in a professional capacity. Conversely, personal interests from Facebook might be more important for social users of Twitter. Another outstanding question is how to carry out detailed quantitative and user-based evaluations of such merged user models, as this has not been discussed by [67]. In addition to these open issues, Kapanipathi *et al* have themselves suggested that future work needs to address also the use of inferencing, based on the richer semantics present in the linked data resources, which are used to ground the automatically derived entities and topics of interest.

Lastly, another challenging question is how to go beyond interest-based models and interaction-based social networks. For instance, Gentile *et al* [49] have demonstrated how people's expertise could be captured from their email exchanges and used to build dynamic user profiles. These are then compared with each other, in order to derive automatically an expertise-based user network, rather than one based on social interactions. Such an approach could be extended and adapted to blogs (e.g. for discovery and recommendation of blogs), as well as to information sharing posts in Twitter and LinkedIn streams.

6. Semantic-based Information Access over Media Streams

Semantic annotations enable users to find documents that mention one or more concepts from the ontology and, optionally, their relations [69]. Depending on the methods used, search queries can often mix free-text keywords with restrictions over semantic annotations (e.g. GATE Mímir [35]). Search tools often provide also browsing functionality, as well as search refinement capabilities [69]. Due to the fact that social media streams are high volume and change over time, semantic search and browsing is a very challenging task.

In general, semantic-based search and retrieval over social media streams differ from traditional information retrieval, due to the additionally available ontological knowledge. On the other hand, they also differ from semantic web search engines, such as Swoogle [40], due to their focus on semantic annotations and using those to retrieve documents, rather than forming queries against ontologies to obtain sets of machine-readable triples.

This section discusses methods specifically developed for social media streams.

6.1. Semantic Search over Social Media Streams

Searching social media streams differs significantly from web searches [137] in a number of important

[12]Estimates suggest that only 25% of tweets contain links: http://techcrunch.com/2010/09/14/twitter-seeing-90-million-tweets-per-day/.

ways. Firstly, users search message streams, such as Twitter, for temporally relevant information and are mostly interested in people. Secondly, searches are used to monitor Twitter content over time and can be saved as part of user profiles. Thirdly, Twitter search queries are significantly shorter and results include more social chatter, whereas web searches look for facts. Coupled with the short message length, noisy nature, and additional information hidden in URLs and hashtags, these differences make traditional keyword-based search methods sub-optimal on media streams. Here we focus on recent work on semantic search, addressing these challenges.

The TREC 2011 Microblog track[13] has given impetus to research by providing a set of query topics, a time point, and a corpus of 16 million tweets, a subset of which was hand-annotated for relevance as a gold standard. In addition to the widely used keyword-based and tweet syntax features (e.g. whether it contains a hashtag), Tao *et al* [136] experimented with entity-based semantic features produced by DBpedia Spotlight, which provided significantly better results.

The Twarql system [86] generates RDF triples from tweets, based on metadata from the tweets themselves, as well as entity mentions, hashtags, and URLs [87]. These are encoded using standard Open Data vocabularies (FOAF, SIOC) (see Section 3) and can be searched through SPARQL queries. It is also possible to subscribe to a stream of tweets matching a complex semantic query, e.g. what competitors are mentioned with my product (Apple iPad in their use case). At the time of writing, Twarql has not been evaluated formally, so its effectiveness and accuracy are yet to be established.

Abel *et al* propose an adaptive faceted search framework for social media streams [1]. It uses semantic entity annotations by OpenCalais, coupled with a user model (see Section 5.1.2), in order to create and rank facets semantically. Keyword search and hashtag-based facets are used as the two baselines. The best results are achieved when facets are personalised, i.e. ranked according to which entities are interesting for the given user (as coded in their entity-based user model). Facet ranking also needs to be made sensitive to the temporal context (essentially the difference between query time and post timestamp).

[13] http://sites.google.com/site/trecmicroblogtrack/

6.2. Filtering and Recommendations for Social Media Streams

The unprecedented rise in the volume and perceived importance of social media content has resulted in individuals starting to experience information overload. In the context of Internet use, research on information overload has shown already that high levels of information can lead to ineffectiveness, as "a person cannot process all communication and informational inputs" [13]. Consequently, researchers are studying information filtering and content recommendation, in order to help alleviate information overload arising from social media streams. Since Facebook streams are predominantly private, the bulk of work has so far focused on Twitter.

As discussed in [31], social media streams are particularly challenging for recommender methods, and different from other types of documents/web content. Firstly, relevance is tightly correlated with recency, i.e. content stops being interesting after just a few days. Secondly, users are active consumers and generators of social content, as well as being highly connected with each other. Thirdly, recommenders need to strike a balance between filtering out noise and supporting serendipity/knowledge discovery. Lastly, interests and preferences vary significantly from user to user, depending on the volume of their personal stream; what and how they use social media for (see Section 5.1.3 on user roles); and user context (e.g. mobile vs tablet, work vs home).

Chen *et al* [31] and Abel *et al* [3] focused on recommending URLs to Twitter users, since it is a common information sharing task. The approach of Chen *et al* is based on a bag-of-words model of user interests, based on the user tweets, what is trending globally, and the user's social network. URL topics are modelled similarly as a word vector and tweet recommendations are computed using cosine similarity.

Abel *et al* [3] improve on this approach by deriving semantic-based user interest models (see Section 5.1.2), which are richer and more generic. They also capture more information through hashtag semantics, replies, and, crucially, by modelling temporal dynamics of user interests.

Recently, Chen *et al* [30] extended their work towards recommending interesting conversations, i.e. threads of multiple messages. The rationale comes from the widespread use of Facebook and Twitter for social conversations [95], coupled with the difficulties that users experience with following these conver-

sations over time, in Twitter in particular. Conversations are rated based on thread length, topic (using bag-of-words as above) and tie-strength (higher priority for content from tightly connected users). Tie strength is modelled for bi-directionally connected users only, using the existence of direct communication, its frequency, and the tie strengths of their mutual friends. Results showed that different recommendation strategies are appropriate for different types of Twitter users, i.e. those who use it for social purposes prefer conversations from closely tied friends, whereas for information seekers, the social connections are much less important.

In the context of Facebook, researchers from Microsoft [97] have trained SVM classifiers to predict, for a given user, the importance of Facebook posts within their news feed, as well as the overall importance of their friends. They also demonstrate a correlation between the two, i.e. the overall importance of a friend influences significantly the importance of posts. In terms of semantic information, the method utilises the Linguistic Inquiry and Word Count (LIWC) dictionary and its 80 topic categories [104]. One of the key findings was the empirical validation of the need for filtering and recommendation of user posts, going beyond reverse chronological order. A second very important, but less strongly substantiated, finding is the need for personalisation (i.e. the same post could be very important for one user, while marked as non-relevant by another).

The issue has recently been recognised by Facebook, who have started to filter the posts shown in the user's news feed, according to the system's proprietary EdgeRank model of importance [68]. EdgeRank takes into account the tie strength (affinity) between the posting user and the viewing user, the type of post (comment, like, etc), and a time decay factor. However, the full details of the algorithm are currently unknown, as is its evaluation. Anecdotally, in 2010 50% of all users were still clicking on the reverse chronological timeline of their feeds. This feature has since been removed and the EdgeRank algorithm refined further. However, it is still not yet possible for the users themselves to train the system, by marking explicitly which posts they consider important.

6.3. Stream Browsing and Visualisation

The main challenge in browsing and visualisation of high-volume stream media is in providing a suitably aggregated, high-level overview. Timestamp-based list interfaces that show the entire, continuously updating stream (e.g. the Twitter timeline-based web interface) are often impractical, especially for analysing high-volume, bursty events. For instance, during the royal wedding in 2011, tweets during the event exceeded 1 million. Similarly, monitoring long running events, such as presidential election campaigns, across different media and geographical locations is equally complex.

One of the simplest and most widely used visualisations is word clouds. These generally use single word terms, which can be somewhat difficult to interpret without extra context. Word clouds have been used to assist users in browsing social media streams, including blog content [11] and tweets [126,96]. For instance, Phelan *et al* [106] use word clouds to present the results of a Twitter based recommendation system. The Eddi system [16] uses topic clouds, showing higher-level themes in the user's tweet stream. These are combined with topic lists, which show who tweeted on which topic, as well as a set of interesting tweets for the highest ranked topics. The Twitris system (see Figure 4) derives even more detailed, contextualised phrases, by using 3-grams, instead of uni-grams [96]. More recently, the concept has been extended towards image clouds [41].

The main drawback of cloud-based visualisations is their static nature. Therefore, they are often combined with timelines showing keyword/topic frequencies over time [4,16,62,141], as well as methods for discovery of unusual popularity bursts [11]. [38] use a timeline which is synchronised with a transcript of a political broadcast, allowing navigation to key points in a video of the event, and displaying tweets from that time period. Overall sentiment is shown on a timeline at each point in the video, using simple colour segments. Similarly, TwitInfo (see Figure 6 [78]) uses a timeline to display tweet activity during a real-world event (e.g. a football game), coupled with some example tweets, colour-coded for sentiment. Some of these visualisations are dynamic, i.e. update as new content comes in (e.g. topic streams [41], falling keyword bars [62] and dynamic information landscapes [62]).

In addition, some visualisations try to capture the semantic relatedness between topics in the media streams. For instance, BlogScope [11] calculates keyword correlations, by approximating mutual information for a pair of keywords using a random sample of documents. Another example is the information landscape visualisation, which conveys topic similarity through spatial proximity [62] (see Figure 5). Topic-

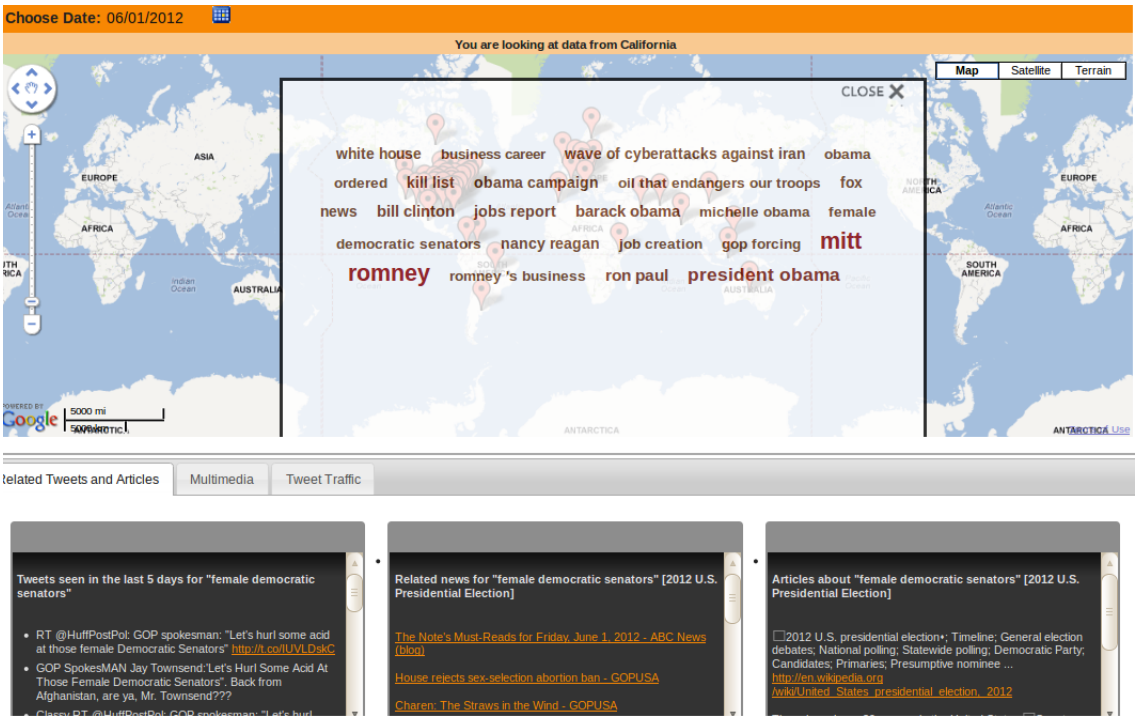

Fig. 4. The Twitris Social Media Event Monitoring Portal (http://twitris.knoesis.org)

Fig. 7. Different Topics Extracted by Twitris for Great Britain

document relationships can be shown also through force-directed, graph-based visualisations [43]. Lastly, Archambault *et al* [7] propose multi-level tag clouds, in order to capture hierarchical relations.

Another important dimension of user-generated content is its place of origin. For instance, some tweets are geo-tagged with latitude/longitude information, while many user profiles on Facebook, Twitter, and blogs specify a user location. Consequently, map-based visualisations of topics have also been explored [88,78,62,96] (see also Figures 5 and 6). For instance, Twitris [96] allows users to select a particular state from the Google map and shows the topics discussed in social media from this state only. Figure 4 shows the Twitris US 2012 Presidential elections monitor, where we have chosen to see the related topics discussed in social media originating from California. Clicking on the topic "female democratic senators" displays the relevant tweets, news, and Wikipedia articles. For comparison, Figure 7 shows the most discussed topics related to the election, extracted from social media originating from Great Britain. While there is significant topic overlap between the two locations, the differences become also clearly visible.

Opinions and sentiment also feature frequently in visual analytics interfaces. For instance, Media Watch (Figure 5 [62]) combines word clouds with aggregated sentiment polarity, where each word is coloured in a shade of red (predominantly negative sentiment), green (predominantly positive), or black (neutral/no sentiment). Search results snippets and faceted browsing terms are also sentiment coloured. Others have combined sentiment-based colour coding with event timelines [4], lists of tweets (Figure 6 [78]), and mood maps [4]. Aggregated sentiment is typically presented using pie charts [141] and, in the case of TwitInfo, the overall statistics are normalised for recall (Figure 6 [78]).

Researchers have also investigated specifically the problem of browsing and visualising social media conversations about real-world events, e.g. broadcast

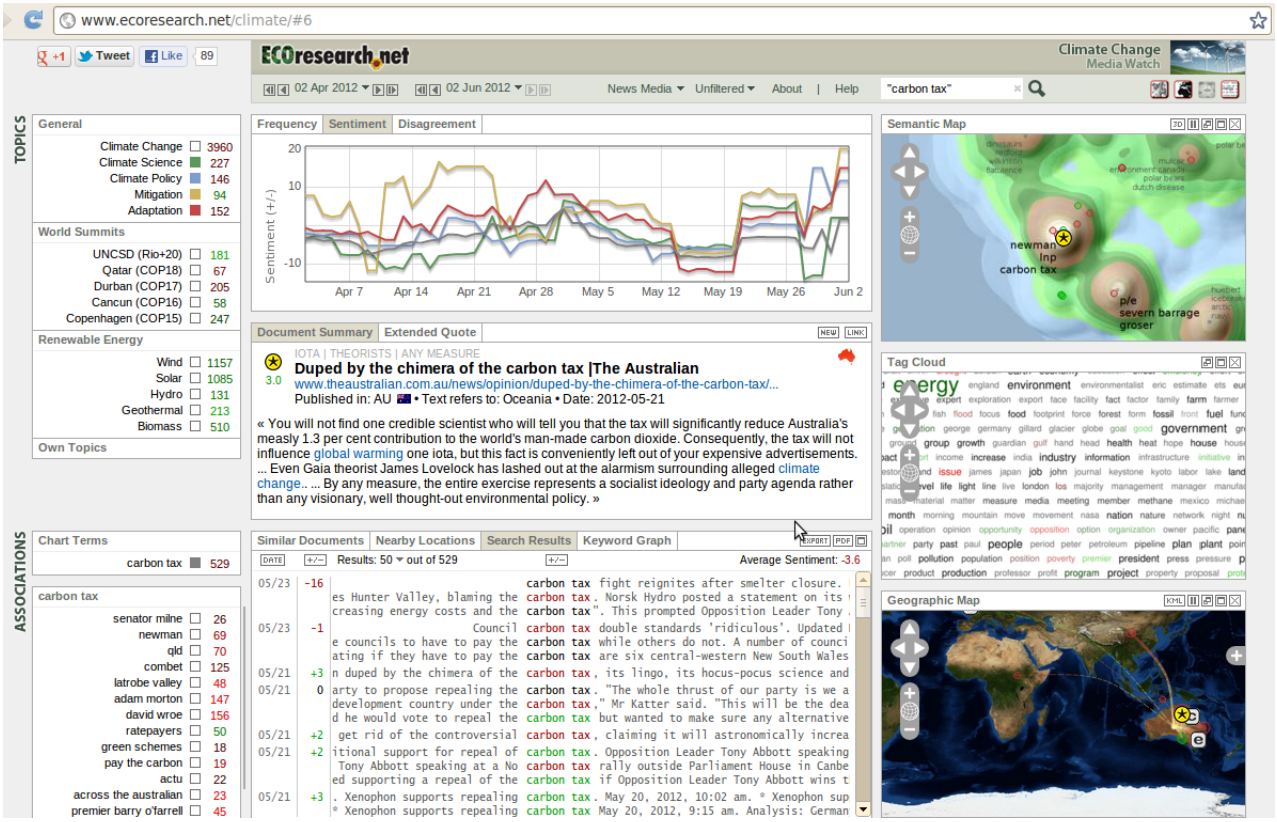

Fig. 5. Media Watch on Climate Change Portal (http://www.ecoresearch.net/climate)

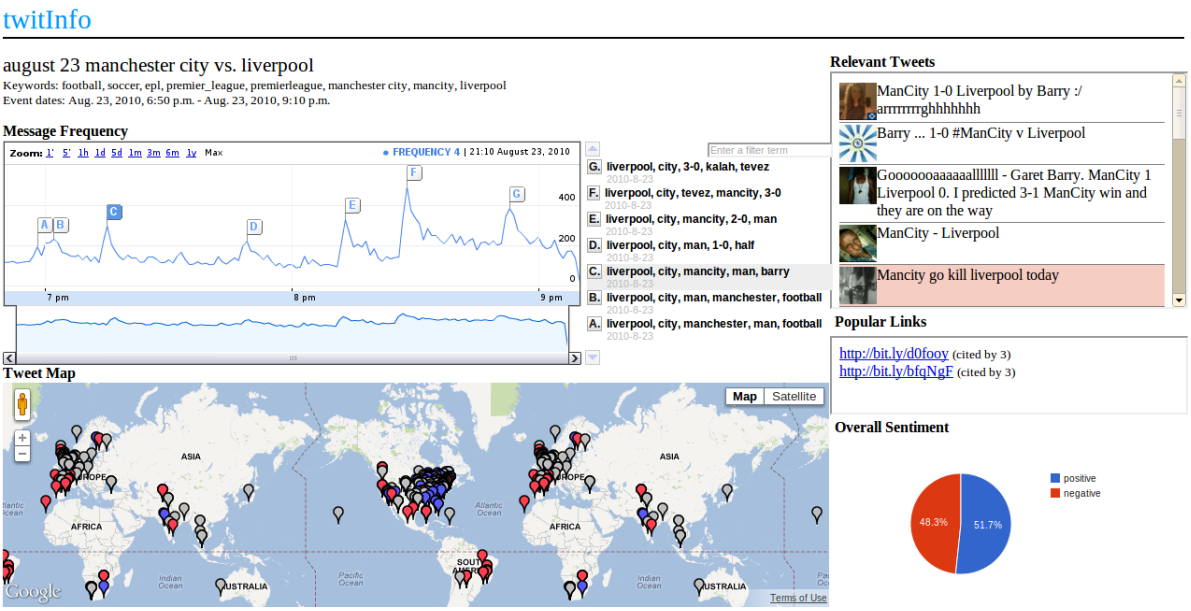

Fig. 6. TwitInfo tracks a football game (http://twitinfo.csail.mit.edu/)

events [126], football games (Figure 6 [78]), conferences [41], and news events [88,4]. A key element here is the ability to identify sub-events and combine these with timelines, maps, and topic-based visualisations.

Lastly, given the user-generated and social nature of the media streams, some visualisations have been designed to exploit this information. For instance, the PeopleSpiral visualisation [41] plots Twitter users who have contributed to a topic (e.g. posted using a given hashtag) on a spiral, starting with the most active and 'original' users first. User originality is measured as the ratio between the number of tweets authored by the user versus re-tweets made. OpinionSpace [46] instead clusters and visualizes users in a two-dimensional space, based on the opinions they have expressed on a given set of topics. Each point in the visualisation shows a user and their comment, so the closer two points, the more similar the users and opinions are. However, the purely point-based visualisation was found hard to interpret by some users, since they could not see the textual content until they clicked on a point. ThemeCrowds [7] instead derives hierarchical clusters of Twitter users through agglomerative clustering and provides a summary of the tweets generated by this user cluster, through multilevel tag clouds (inspired by treemap visualisation). Tweet volumes over time are shown in a timeline-like view, which also allows the selection of a time period.

6.4. Discussion

Most current search, recommendation, and visualisation methods tend to use shallow textual and frequency-based information. For instance, a comparison between TF-IDF weighted topic models and LDA topic modelling has shown the former to be superior [30,114]. However, these can be improved further through integration of semantic information, as suggested by [30]. In the case of personalised recommendations, these could be improved by incorporating user behaviour roles, making better use of the latent semantics and implicit user information, as well as better integration of the temporal dimension in the recommender algorithms.

Browsing and visualisation interfaces can also be improved by taking into account the extra semantic knowledge about the entities mentioned in the media streams. For instance, when entities and topics are annotated with URIs to LOD resources, such as DBpedia, the underlying ontology can underpin hierarchically-based visualisations, including semantic relations. In addition, the exploration of media streams through topic-, entity-, and time-based visualisations can be enriched with ontology-based faceted search and semantic query interfaces. One such example is the KIM semantic platform, which is, however, aimed at largely static document collections [110].

Algorithm scalability and efficiency are particularly important, due to the large-scale, dynamic nature of social media streams. For instance, the interactive Topic Stream visualisation takes 45 seconds to compute on 1 million tweets and 325,000 contributing users, which is too long for most usage scenarios [41]. Similarly, calculating keyword correlations through point-wise mutual information is computationally too expensive on high volume blog posts [11]. A frequently used solution is to introduce a sliding window over the data (e.g. between one week and one year) and thus limit the content used for IDF and other such calculations.

In conclusion, designing effective semantic search, browsing and visualisations interfaces for media streams has proven particularly challenging. Based on our survey of the state-of-the-art, we have derived the following requirements:

- designing meaningful and intuitive visualisations, conveying intuitively the complex, multi-dimensional semantics of user-generated content (e.g. topics, entities, events, user demographics (including geolocation), sentiment, social networks);
- visualising changes over time;
- supporting different levels of granularity, both at the level of semantic content, user clusters, and temporal windows;
- allowing interactive, real-time exploration;
- integration with search, to allow users to select a subset of relevant content;
- exposing the discussion/threaded nature of the social conversations;
- addressing scalability and efficiency.

Amongst the systems surveyed, only Twitris [96] and Media Watch [62] have started to address most of these requirements, but not without limitations. Firstly, their current visualisations are mostly topic- and entity-centric and could benefit from integration of event-based visualisations, such as TwitInfo [78] and Tweetgeist [126]. Secondly, user demographics as means for stream media aggregation and exploration is mostly limited to map-based visualisations. Additional search and browsing capabilities, based around users' age, gender, political views, interests, and other such

characteristics are also needed. Thirdly, methods for information aggregation and exploration, based on social networks (e.g hubs and authorities) could be combined with the currently prevailing topic- and content-centric approaches. Lastly, we would like to advocate a more substantial end-user involvement in the design and testing of new intelligent information access systems. In this way, the resulting user interfaces will address the emerging complex information seeking requirements, in terms of better support for sense making, learning and investigation, and social search [107].

7. Outstanding Challenges and Conclusions

This paper set out to explore a number of research questions arising from applications of semantic technologies to social media.

Firstly, we examined existing ontologies in the context of modelling the semantics of social media streams. Our conclusion is that most applications tend to adopt or extend more than one ontology, since they model different aspects. With respect to Web of Data resources, current methods have made most use of Wikipedia-derived resources (namely DBpedia and YAGO) and, to a lesser degree – Geonames, Freebase, and domain-specific ones like MusicBrainz. Better exploiting this wealth of semantic knowledge for semantic annotation of social media remains a challenge, which we discussed in more detail in section 4.6.

Next, the questions of capturing the implicit semantics and dealing with the noisy, dynamic nature of social media streams, were addressed as part of our analysis of semantic annotation state-of-the-art. We identified the need for more robust and accurate large-scale entity and event recognition methods, as well as finer-grained opinion mining algorithms to address target identification, volatility over time, detecting and modelling conflicting opinions, and opinion aggregation (see section 4.6 for details).

Thirdly, current methods for modelling users' digital identity and social media activities were discussed. Limitations with respect to modelling user interests and integration of temporal dynamics were identified, coupled with emerging need for cross-media user models. A more in-depth discussion appears in section 5.2.

Lastly, semantic-based methods for search, browsing, recommendation, and information visualisation of social media content were reviewed, from the perspective of supporting complex information seeking behaviour. As a result, seven key requirements were identified and limitations of current approaches were discussed in this context.

In conclusion, we discuss three major areas where further research is necessary.

7.1. Cross-Media Aggregation and Multilinguality

The majority of methods surveyed here have been developed and evaluated only on one kind of social media (e.g. Twitter or blog posts). Cross-media linking, going beyond connecting tweets to news articles, is a crucial open issue, due to the fact that increasingly users are adopting more than one social media platform, often for different purposes (e.g. personal vs professional use). In addition, as people's lives are becoming increasingly digital, this work will also provide a partial answer to the challenge of inter-linking our personal collections (e.g. emails, photos) with our social media online identities.

The challenge is to build computational models of cross-media content merging, analysis, and visualisation and embed these into algorithms capable of dealing with the large-scale, contradictory and multi-purpose nature of multi-platform social media streams. For example, further work is needed on algorithms for cross-media content clustering, cross-media identity tracking, modelling contradictions between different sources, and inferring change in interests and attitudes over time.

Another related major challenge is multilinguality. Most of the methods surveyed here were developed and tested on English content only. As discussed in Section 4.6, some initial steps are being made through multilingual lexicons, such as Wiktionary [89] and UBY [55], and linguistically grounded ontologies [22]. Other work has focused on widening the range of available linguistic resources to less studied languages, through crowd-sourcing. Amazon Mechanical Turk, in particular, has emerged as particularly useful, since crowd-sourcing projects are easily setup there, coupled with the fact that it allows "access to foreign markets with native speakers of many rare languages" [145]. This feature is particularly useful for researchers working on less-resourced languages, such as Arabic [45], Urdu [145] and others [5,24,65]. Irvine and Klementiev [65], for example, have shown that it is possible to create lexicons between English and 37 out of the 42 low resource languages that they experimented with. Similarly, Weichselbraun *et al* [140] crowd-source domain-specific sentiment lexicons in multiple languages, through games with a purpose. A

related aspect is designing crowdsourcing projects, so that they can be re-used easily across languages, e.g. [65,76] for Mechanical Turk and [109,113] for games-with-a-purpose. There is also the related issue of annotated corpora and evaluation, to which we return in Section 7.3 below.

Lastly, as users are increasingly consuming social media streams on different hardware platforms (desktops, tablets, smart phones), cross-platform and/or platform-independent information access methods need to be developed. This is particularly challenging in the case of information visualisation on small screen devices.

7.2. Scalability and Robustness

In information extraction research, large-scale algorithms (also referred to as data-intensive or web-scale natural language processing) are demonstrating increasingly superior results compared to approaches trained on smaller datasets [56]. This is mostly thanks to addressing the data sparseness issue through collection of significantly larger numbers of naturally occurring linguistic examples [56]. The need for and the success of data-driven NLP methods to a large extent mirrors recent trends in other research fields, leading to what is being referred to as "the fourth paradigm of science" [12].

At the same time, semantic annotation and information access algorithms need to be scalable and robust, also in order to cope with the large content volumes encountered in social media streams. Many use cases require online, near real-time processing, which introduces additional requirements in terms of algorithm complexity. Cloud computing [39] is increasingly being regarded as a key enabler of scalable, on-demand processing, giving researchers everywhere affordable access to computing infrastructures, which allow the deployment of significant compute power on an on-demand basis, and with no upfront costs.

However, developing scalable and parallelisable algorithms for platforms such as Hadoop is far from trivial. Straightforward deployment and sharing of semantic annotation pipelines and algorithm parallelisation are only a few of the requirements which need to be met. Research in this area is still in its infancy, especially around general purpose platforms for scalable semantic processing.

GateCloud.net [134] can be viewed as the first step in this direction. It is a novel cloud-based platform for large-scale text mining research, which also supports ontology-based semantic annotation pipelines. It aims to provide researchers with a platform-as-a-service, which enables them to carry out large-scale NLP experiments by harnessing the vast, on-demand compute power of the Amazon cloud. It also minimises the need to implement specialised parallelisable text processing algorithms. Important infrastructural issues are dealt with by the platform, completely transparently for the researcher: load balancing, efficient data upload and storage, deployment on the virtual machines, security, and fault tolerance.

One example application of GateCloud was in a project with the UK National Archive [80], which used it to annotate semantically 42 TB of web pages and other textual content. The semantic annotation process was underpinned by a large-scale knowledge base, acquired from the LOD cloud, data.gov.uk, and a large geographical database. The results were indexed in GATE Mímir [35], coupled with a user interface for browsing, search and navigation from the document space into the semantic knowledge base via full-text search, semantic annotations and SPARQL queries.

7.3. Evaluation, Shared Datasets and Crowdsourcing

The third major open issue is evaluation. As discussed in all three application areas of semantic technologies for social media streams, lack of shared gold-standard datasets is hampering repeatability and comparative evaluation of algorithms. At the same time, comprehensive user- and task-based evaluation experiments are also required, in order to identify problems with existing search and visualisation methods. Particularly in the area of intelligent information access, most of the papers surveyed either did not report evaluation experiments, or those that did, tended to carry out small-scale, formative studies. Longitudinal evaluation with larger user groups is particularly lacking.

Similarly, algorithm training and adaptation on social media gold standard datasets is currently very limited. For example, no gold standard datasets of Twitter and blog summaries exist and there are fewer than 10,000 tweets annotated with named entities. Creating sufficiently large, vitally needed datasets through traditional expert-based text annotation methodologies is very expensive, both in terms of time and funding required. The latter can vary between USD 0.36 and 1.0 [109], which is unaffordable for corpora consisting of millions of words. Some cost reductions could be achieved through web-based collaborative annotation tools, such as GATE Teamware [21], which sup-

port distributed teams and are tailored to non-expert annotators.

Another alternative are commercial crowdsourcing marketplaces have been reported to be 33% less expensive than in-house employees on tasks such as tagging and classification [61]. Consequently, in the field of language processing, researchers have started creating annotated corpora with Amazon Mechanical Turk and game-based approaches as less expensive alternatives. A comprehensive survey of crowdsourcing is beyond the scope of this paper, however, for details see [120,25,112].

With respect to corpus annotation in particular, Poesio *et al* [109] estimate that, compared to the cost of expert-based annotation (estimated as $1.000.000), the cost of 1 million annotated tokens could be indeed reduced to less than 50% by using MTurk (i.e., $380.000 - $430.000) and to around 20% (i.e., $217,927) when using a game based approach such as their own PhraseDetectives game. With respect to crowdsourcing social media annotations, there have been experiments on, e.g., categorizing tweets [102] and annotating named entities in tweets [47]. In the Semantic Web field, researchers have explored mostly crowdsourcing through games with a purpose, primarily for knowledge acquisition [129,138] and LOD improvement [139].

At the same time, researchers have turned to crowdsourcing as a means for scaling up human-based evaluation experiments. The main challenge here is in how to define the evaluation task, so that it can be crowdsourced from non-specialists, with high quality results [82].This is far from trivial, and researchers have argued that crowdsourcing evaluation tasks need to be designed differently from expert-based evaluations [50]. In particular, Gillick and Liu [50] found that non-expert evaluation of summarization systems produces noisier results, thus requiring more redundancy to achieve statistical significance and that Mechanical Turk workers cannot produce score rankings that agree with expert ranking.

One successful design for crowdsourcing-based evaluation has used a four phase workflow of separate tasks, which has been tried on reading comprehension of machine translation [24]. Another, simpler task design has been used by [63] for evaluation of tweet summaries. Inouye *et al* asked Mechanical Turk workers to indicate on a five point scale, how much of the information from the human produced summary is contained in the automatically produced summary. A third evaluation example, which has achieved successful results on Mechanical Turk, is pair-wise ranking [51]. The task in this case is to identify the most informative sentence from a product review. In this case, the crowdworkers are asked to indicate whether a sentence chosen by the baseline system is more informative than a sentence chosen by the author's method. Sentence order is randomised and it is also possible to indicate that none of these sentences are a good summary.

To conclude, crowdsourcing has recently emerged as a promising method for creating shared evaluation datasets, as well as for carrying out user-based evaluation experiments. Adapting these efforts to the specifics of semantic annotation and information visualisation, as well as using these to create large-scale resources and repeatable, longitudinal evaluations, are key areas for future work.

Acknowledgements

This work was supported by funding from the Engineering and Physical Sciences Research Council (grant EP/I004327/1). The authors wish to thank Marta Sabou and Arno Scharl for the discussions on crowdsourcing and its role in semantic technologies research, as well as Diana Maynard for the discussions on opinion mining of Twitter and Facebook messages and for proof-reading the paper. We would also like to thank the reviewers and the editor for their comments and suggestions, which helped us improve this paper significantly.

References

[1] F. Abel, I. Celik, G.-J. Houben, and P. Siehndel. Leveraging the semantics of tweets for adaptive faceted search on Twitter. In *Proceedings of the 10th international conference on The semantic web - Volume Part I*, ISWC'11, pages 1–17, Berlin, Heidelberg, 2011. Springer-Verlag.

[2] F. Abel, Q. Gao, G. J. Houben, and K. Tao. Semantic enrichment of Twitter posts for user profile construction on the social web. In *ESWC (2)*, pages 375–389, 2011.

[3] F. Abel, Q. Gao, G.J. Houben, and K.ele Tao. Analyzing temporal dynamics in Twitter profiles for personalized recommendations in the social web. In *Proceedings of 3rd International Conference on Web Science (WebSci'11)*, Koblenz, Germany, 2011.

[4] B. Adams, D. Phung, and S. Venkatesh. Eventscapes: Visualizing events over time with emotive facets. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 1477–1480, 2011.

[5] Vamshi Ambati and Stephan Vogel. Can crowds build parallel corpora for machine translation systems? In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 62–65, 2010.

[6] S. Angeletou, M. Rowe, and H. Alani. Modelling and analysis of user behaviour in online communities. In *Proceedings of the 10th International Conference on the Semantic Web*, ISWC'11, pages 35–50. Springer-Verlag, 2011.

[7] D. Archambault, D. Greene, P. Cunningham, and N. J. Hurley. ThemeCrowds: Multiresolution summaries of Twitter usage. In *Workshop on Search and Mining User-Generated Contents (SMUC)*, pages 77–84, 2011.

[8] L. Aroyo and G.-J. Houben. User modeling and adaptive semantic web. *Semantic Web*, 1(1,2):105–110, April 2010.

[9] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th International Conference on World Wide Web*, pages 61–70. ACM, 2010.

[10] T. Baldwin and M. Lui. Language identification: The long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 229–237, Los Angeles, California, June 2010.

[11] N. Bansal and N. Koudas. Blogscope: Spatio-temporal analysis of the blogosphere. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 1269–1270, 2007.

[12] Roger Barga, Dennis Gannon, and Daniel Reed. The client and the cloud: Democratizing research computing. *IEEE Internet Computing*, 15(1):72–75, 2011.

[13] C. Beaudoin. Explaining the relationship between internet use and interpersonal trust: Taking into account motivation and information overload. *Journal of Computer Mediated Communication*, 13:550—568, 2008.

[14] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *Proceedings of the Third International Conference on Web Search and Web Data Mining*, pages 291–300, 2010.

[15] Hila Becker, Mor Naaman, and Luis Gravano. Beyond Trending Topics: Real-World Event Identification on Twitter. In *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM)*, 2011.

[16] 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 *Proceedings of the 23nd ACM Symposium on User Interface Software and Technology (UIST)*, pages 303–312, 2010.

[17] Toine Bogers and Antal van den Bosch. Fusing recommendations for social bookmarking websites. *International Journal of Electronic Commerce*, 15(3):33–75, Spring 2011.

[18] E. Boiy and M-F. Moens. A machine learning approach to sentiment analysis in multilingual web texts. *Information Retrieval*, 12(5):526–558, 2009.

[19] J. Bollen, A. Pepe, and H. Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *http://arxiv.org/abs/0911.1583*, 2009.

[20] K. Bontcheva and H. Cunningham. Semantic annotation and retrieval: Manual, semi-automatic and automatic generation. In J. Domingue, D. Fensel, and J. A. Hendler, editors, *Handbook of Semantic Web Technologies*. Springer, 2011.

[21] Kaling Bontcheva, Hamish Cunningham, Ian Roberts, Angus Roberts, Valentin Tablan, Niraj Aswani, and Genevieve Gorrell. GATE Teamware: A Web-based, Collaborative Text Annotation Framework. *Language Resources and Evaluation*, Forthcoming.

[22] Paul Buitelaar, Philipp Cimiano, Peter Haase, and Michael Sintek. Towards Linguistically Grounded Ontologies. In *Proceedings of the European Semantic Web Conference (ESWC'09), LNCS 5554*, pages 111—125, 2009.

[23] J. Burger, J. Henderson, G. Kim, and G. Zarrella. Discriminating Gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1301–1309, 2011.

[24] Chris Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, 2009.

[25] Chris Callison-Burch and Mark Dredze. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12, 2010.

[26] S. Carter, W. Weerkamp, and E. Tsagkias. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation Journal*, Forthcoming.

[27] Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic grounding of tag relatedness in social bookmarking systems. In *Proceedings of the 7th International Conference on The Semantic Web*, pages 615–631, 2008.

[28] I. Celino, D. Dell'Aglio, E. Della Valle, Y. Huang, T. Lee, S. Park, and V. Tresp. Making Sense of Location-based Micro-posts Using Stream Reasoning. In *Proceedings of the Making Sense of Microposts Workshop (#MSM2011), collocated with the 8th Extended Semantic Web Conference, Heraklion, Crete, Greece*, 2011.

[29] J. Chan, C. Hayes, and E. Daly. Decomposing discussion forums using common user roles. In *Proceedings of WebSci10: Extending the Frontiers of Society On-Line*, 2010.

[30] J. Chen, R. Nairn, and E. Chi. Speak little and well: Recommending conversations in online social streams. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*, CHI '11, pages 217–226, 2011.

[31] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: Experiments on recommending content from information streams. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, CHI '10, pages 1185–1194, 2010.

[32] Y. Cheng, G. Qiu, J. Bu, K. Liu, Y. Han, C. Wang, and C. Chen. Model bloggers' interests based on forgetting mechanism. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 1129–1130, 2008.

[33] Z. Cheng. You are where you tweet: A content-based approach to geo-locating Twitter users. *Proceedings of the 19th ACM Conference*, 2010.

[34] S. Choudhury and J. Breslin. Extracting semantic entities and events from sports tweets. In *Proceedings, 1st Workshop on Making Sense of Microposts (#MSM2011): Big things come in small packages*, pages 22–32, 2011.

[35] Hamish Cunningham, Valentin Tablan, Ian Roberts, Mark A. Greenwood, and Niraj Aswani. Information Extraction and Semantic Annotation for Multi-Paradigm Information Management. In Mihai Lupu, Katja Mayer, John Tait, and Anthony J. Trippe, editors, *Current Challenges in Patent Information Retrieval*, volume 29 of *The Information Retrieval Series*. Springer, 2011.

[36] S. Decker and M. Frank. The Social Semantic Desktop. Technical report, DERI Technical Report 2004-05-02, 2004.

[37] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. ZenCrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st Conference on World Wide Web*, pages 469–478, 2012.

[38] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, pages 115–122, 2010.

[39] Marios D Dikaiakos, Dimitrios Katsaros, Pankaj Mehra, George Pallis, and Athena Vakali. Cloud computing: Distributed internet computing for IT and scientific research. *IEEE Internet Computing*, 13(5):10–13, 2009.

[40] Li Ding, Tim Finin, Anupam Joshi, Rong Pan, R. Scott Cost, Yun Peng, Pavan Reddivari, Vishal C. Doshi, and Joel Sachs. Swoogle: A Search and Metadata Engine for the Semantic Web. In *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*, 2004.

[41] M. Dork, D. Gruen, C. Williamson, and S. Carpendale. A visual backchannel for large-scale events. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1129–1138, November 2010.

[42] M. Dowman, V. Tablan, H. Cunningham, and B. Popov. Web-assisted annotation, semantic indexing and search of television and radio news. In *Proceedings of the 14th International World Wide Web Conference*, Chiba, Japan, 2005.

[43] J. Eisenstein, D. H. P. Chau, A. Kittur, and E. Xing. Topicviz: Semantic navigation of document collections. In *CHI Work-in-Progress Paper (Supplemental Proceedings)*, 2012.

[44] J. Eisenstein, B. O'Connor, N.A. Smith, and E.P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, 2010.

[45] Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. Using Mechanical Turk to create a corpus of Arabic summaries. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, 2010.

[46] S. Faridani, E. Bitton, K. Ryokai, and K. Goldberg. Opinion space: A scalable tool for browsing online comments. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI)*, pages 1175–1184, 2010.

[47] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88, 2010.

[48] Clay Fink, Christine Piatko, James Mayfield, Tim Finin, and Justin Martineau. Geolocating blogs from their textual content. In *Working Notes of the AAAI Spring Symposium on Social Semantic Web: Where Web 2.0 Meets Web 3.0*, pages 1–2. AAAI Press, 2008.

[49] A. L. Gentile, V. Lanfranchi, S. Mazumdar, and F. Ciravegna. Extracting semantic user networks from informal communication exchanges. In *Proceedings of the 10th International Conference on the Semantic Web*, ISWC'11, pages 209–224. Springer-Verlag, 2011.

[50] Dan Gillick and Yang Liu. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151, 2010.

[51] Andrea Glaser and Hinrich Schütze. Automatic generation of short informative sentiment summaries. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 276–285, Avignon, France, April 2012.

[52] A. Go, R. Bhayani, , and L. Huang. Twitter sentiment classification using distant supervision. Technical Report CS224N Project Report, Stanford University, 2009.

[53] S. Gouws, D. Metzler, C. Cai, and E. Hovy. Contextual bearing on linguistic variation in social media. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 20–29, 2011.

[54] D. Gruhl, M. Nagarajan, J. Pieper, C. Robson, and A. Sheth. Context and Domain Knowledge Enhanced Entity Spotting in Informal Text. In *Proceedings of the 8th International Semantic Web Conference (ISWC'2009)*, 2009.

[55] Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. UBY: A large-scale unified lexical-semantic resource based on LMF. In *13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590, 2012.

[56] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.

[57] B. Han and T. Baldwin. Lexical normalisation of short text messages: makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 368–378, 2011.

[58] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, New York, NY, USA, 2011.

[59] Dominik Heckmann, Tim Schwartz, Boris Brandherm, Michael Schmitz, and Margeritta von Wilamowitz-Moellendorff. GUMO - the general user model ontology. In *Proceedings of the 10th International Conference on User Modeling*, pages 428–432, 2005.

[60] M. Hepp. HyperTwitter: Collaborative knowledge engineering via Twitter messages. In *Knowledge Engineering and Management by the Masses - 17th International Conference EKAW 2010*, pages 451–461, 2010.

[61] Leah Hoffmann. Crowd control. *Communications of the ACM*, 52(3):16 –17, 2009.

[62] A. Hubmann-Haidvogel, A. M. P. Brasoveanu, A. Scharl, M. Sabou, and S. Gindl. Visualizing contextual and dynamic features of micropost streams. In *Proceedings of the #MSM2012 Workshop, CEUR*, volume 838, 2012.

[63] David Inouye and Jugal K. Kalita. Comparing Twitter summarization algorithms for multiple post summaries. In *Social-*

Com/PASSAT, pages 298–306, 2011.

[64] N. Ireson and F. Ciravegna. Toponym resolution in social media. In *Proceedings of the 9th International Semantic Web Conference (ISWC)*, pages 370–385, 2010.

[65] Ann Irvine and Alexandre Klementiev. Using Mechanical Turk to annotate lexicons for less commonly used languages. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 108–113, 2010.

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

[67] P. Kapanipathi, F. Orlandi, A. Sheth, and A. Passant. Personalized Filtering of the Twitter Stream. In *2nd workshop on Semantic Personalized Information Management at ISWC 2011*, 2011.

[68] J. Kincaid. EdgeRank: The secret sauce that makes Facebook's news feed tick, April 2010.

[69] A. Kiryakov, B. Popov, D. Ognyanoff, D. Manov, A. Kirilov, and M. Goranov. Semantic annotation, indexing and retrieval. *Journal of Web Semantics, ISWC 2003 Special Issue*, 1(2):671–680, 2004.

[70] Christian Körner, Dominik Benz, Andreas Hotho, Markus Strohmaier, and Gerd Stumme. Stop thinking, start tagging: Tag semantics emerge from collaborative verbosity. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*, pages 521–530, 2010.

[71] Patrick Lai. Extracting Strong Sentiment Trends from Twitter. *http://nlp.stanford.edu/courses/cs224n/2011/reports/patlai.pdf*, 2010.

[72] David Laniado and Peter Mika. Making sense of Twitter. In *International Semantic Web Conference (1)*, pages 470–485, 2010.

[73] Y. Li, K. Bontcheva, and H. Cunningham. Hierarchical, Perceptron-like Learning for Ontology Based Information Extraction. In *16th International World Wide Web Conference (WWW2007)*, pages 777–786, May 2007.

[74] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences*, 102(33):11623–11628, August 2005.

[75] X. Liu, S. Zhang, F. Wei, and M. Zhou. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 359–367, 2011.

[76] Nitin Madnani, Jordan Boyd-Graber, and Philip Resnik. Measuring transitivity using untrained annotators. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 188–194, 2010.

[77] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Where is this tweet from? Inferring home locations of Twitter users. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, Dublin, Ireland, 2012.

[78] 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 *Proceedings of the 2011 Conference on Human Factors in Computing Systems (CHI)*, pages 227–236, 2011.

[79] D. Maynard and A. Funk. Automatic detection of political opinions in tweets. In Dieter Fensel Raúl García-Castro and Grigoris Antoniou, editors, *The Semantic Web: ESWC 2011 Selected Workshop Papers, Lecture Notes in Computer Science*. Springer, 2011.

[80] D. Maynard and M. A. Greenwood. Large Scale Semantic Annotation, Indexing and Search at The National Archives. In *Proceedings of LREC 2012*, Turkey, 2012.

[81] Diana Maynard, Kalina Bontcheva, and Dominic Rout. Challenges in developing opinion mining tools for social media. In *Proceedings of @NLP can u tag #usergeneratedcontent?! Workshop at LREC 2012*, Turkey, 2012.

[82] Richard McCreadie, Craig Macdonald, and Iadh Ounis. Identifying top news using crowdsourcing. *Information Retrieval*, pages 1–31, 2012. 10.1007/s10791-012-9186-z.

[83] L. K. McDowell and M. Cafarella. Ontology-Driven Information Extraction with OntoSyphon. In *5th Internal Semantic Web Conference (ISWC'06)*. Springer, 2006.

[84] E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *Proc. of the Fifth Int. Conf. on Web Search and Data Mining (WSDM)*, 2012.

[85] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia Spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8, 2011.

[86] P. N. Mendes, A. Passant, and P. Kapanipathi. Twarql: Tapping into the wisdom of the crowd. In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 45:1–45:3, 2010.

[87] P. N. Mendes, A. Passant, P. Kapanipathi, and A. P. Sheth. Linked open social signals. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, WI-IAT '10, pages 224–231, Washington, DC, USA, 2010. IEEE Computer Society.

[88] B. Meyer, K. Bryan, Y. Santos, and B. Kim. TwitterReporter: Breaking news detection and visualization through the geotagged Twitter network. In *Proceedings of the ISCA 26th International Conference on Computers and Their Applications*, pages 84–89, 2011.

[89] Christian M. Meyer and Iryna Gurevych. Wiktionary: a new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In *Electronic Lexicography*. Oxford University Press, 2012.

[90] M. Michelson and S. A. Macskassy. Discovering users' topics of interest on Twitter: A first look. In *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*, AND '10, pages 73–80, 2010.

[91] R. Mihalcea and P. Tarau. TextRank: Bringing order into text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 404–411, 2004.

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

[93] D. Milne and I. H. Witten. Learning to link with Wikipedia. In *Proc. of the 17th Conf. on Information and Knowledge Management (CIKM)*, pages 509–518, 2008.

[94] G. Mishne. AutoTag: A collaborative approach to automated tag assignment for weblog posts. In *Proceedings of the 15th International Conference on World Wide Web*, pages 953–954, 2006.

[95] M. Naaman, J. Boase, and C. Lai. Is it really about me?: Message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer Supported Cooperative Work*, pages 189–192. ACM, 2010.

[96] M. Nagarajan, K. Gomadam, A. Sheth, A. Ranabahu, R. Mutharaju, and A. Jadhav. Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences. In *Web Information Systems Engineering*, pages 539–553, 2009.

[97] T. Paek, M. Gamon, S. Counts, D. M. Chickering, and A. Dhesi. Predicting the importance of newsfeed posts and social network friends. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, 2010.

[98] A. Pak and P. Paroubek. Twitter Based System: Using Twitter for Disambiguating Sentiment Ambiguous Adjectives. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 436–439, 2010.

[99] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Information Retrieval*, 2(1), 2008.

[100] A. Passant, J. G. Breslin, and S. Decker. Rethinking microblogging: open, distributed, semantic. In *Proceedings of the 10th International Conference on Web Engineering*, pages 263–277, 2010.

[101] A. Passant and P. Laublet. Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. In *Proceedings of the Linked Data on the Web Workshop (LDOW), Beijing, China*, 2008.

[102] Sharoda A. Paul, Lichan Hong, and Ed H. Chi. What is a question? crowdsourcing tweet categorization. In *CHI'2011 Worshop on Crowdsourcing and Human Computation*, 2011.

[103] M. Pennacchiotti and A.M. Popescu. A Machine Learning Approach to Twitter User Classification. In *Proceedings of ICWSM 2011*, pages 281–288, 2011.

[104] J. D Pennebaker, C. K. Chung, M. Ireland, Gonzales A., and R. J. Booth. The LIWC2007 Application. Technical report, 2007. http://www.liwc.net/liwcdescription.php.

[105] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to Twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189, 2010.

[106] O. Phelan, K. McCarthy, and B. Smyth. Using Twitter to recommend real-time topical news. In *Proceedings of the 2009 ACM Conference on Recommender Systems*, pages 385–388, 2009.

[107] Peter Pirolli. Powers of 10: Modeling complex information-seeking systems at multiple scales. *IEEE Computer*, 42(3):33–40, 2009.

[108] T. Plumbaum, S. Wu, E. W. De Luca, and S. Albayrak. User Modeling for the Social Semantic Web. In *2nd Workshop on Semantic Personalized Information Management: Retrieval and Recommendation, in conjunction with ISWC 2011*, 2011.

[109] M. Poesio, U. Kruschwitz, J. Chamberlain, L. Robaldo, and L. Ducceschi. Phrase Detectives: Utilizing Collective Intelligence for Internet-Scale Language Resource Creation. *Transactions on Interactive Intelligent Systems*, 2012. To Appear.

[110] Borislav Popov, Atanas Kiryakov, Angel Kirilov, Dimitar Manov, Damyan Ognyanoff, and Miroslav Goranov. KIM – Semantic Annotation Platform. In *2nd International Semantic Web Conference (ISWC2003)*, pages 484–499, Berlin, 2003. Springer.

[111] L. Qu, C. Müller, and I. Gurevych. Using tag semantic network for keyphrase extraction in blogs. In *Proceedings of the 17th Conference on Information and Knowledge Management*, pages 1381–1382, 2008.

[112] Alexander J. Quinn and Benjamin B. Bederson. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*, CHI '11, pages 1403–1412, 2011.

[113] W. Rafelsberger and A. Scharl. Games with a purpose for social networking platforms. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, HT '09, pages 193–198, 2009.

[114] Daniel Ramage, Susan Dumais, and Dan Liebling. Characterizing microblogs with topic models. In *Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM)*, 2010.

[115] N. Ravikant and A. Rifkin. Why Twitter Is Massively Undervalued Compared To Facebook. *TechCrunch*, 2010. http://techcrunch.com/2010/10/16/why-twitter-is-massively-undervalued-compared-to-facebook/.

[116] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In *Proc. of Empirical Methods for Natural Language Processing (EMNLP)*, Edinburgh, UK, 2011.

[117] G. Rizzo, R. Troncy, S. Hellmann, and M. Bruemmer. NERD meets NIF: Lifting NLP extraction results to the Linked Data cloud. In *5th Workshop on Linked Data on the Web (LDoW)*, 2012.

[118] M. Rowe, S. Angeletou, and H. Alani. Predicting discussions on the social semantic web. In *Proceedings of the 8th Extended Semantic Web Conference on the Semanic Web*, ESWC'11, pages 405–420. Springer-Verlag, 2011.

[119] M. Rowe and M. Stankovic. Aligning Tweets with Events : Automation via Semantics. *Semantic Web*, 1, 2009.

[120] Marta Sabou, Kalina Bontcheva, and Arno Scharl. Crowdsourcing research opportunities: Lessons from natural language processing. In *12th International Conference on Knowledge Management and Knowledge Technologies (I-KNOW)*, pages 17–24, 2012.

[121] H. Saif, Y. He, and H. Alani. Alleviating data sparsity for Twitter sentiment analysis. In *Proceedings of the #MSM2012 Workshop, CEUR*, volume 838, 2012.

[122] Takeshi Sakaki, M Okazaki, and Y Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web (WWW)*, pages 851–860. ACM, 2010.

[123] H Sayyadi, M Hurst, and A Maykov. Event Detection and Tracking in Social Streams. In *Proceedings of the Third International ICWSM Conference*, pages 311–314, 2009.

[124] S. Scerri, K. Cortis, I. Rivera, and S. Handschuh. Knowledge Discovery in Distributed Social Web Sharing Activities. In *Proceedings of the #MSM2012 Workshop, CEUR*, volume 838, 2012.

[125] A. Scharl and A. Weichselbraun. An automated approach to investigating the online media coverage of US presidential elections. *Journal of Information Technology and Politics*, 5(1):121–132, 2008.

[126] D.A. Shamma, L. Kennedy, and E.F. Churchill. Tweetgeist: Can the Twitter timeline reveal the structure of broadcast events? In *Proceedings of CSCW 2010*, 2010.

[127] B. Sharifi, M. A. Hutton, and J. Kalita. Summarizing Microblogs Automatically. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 685–688, Los Angeles, California, June 2010.

[128] W. Shen, J. Wang, P. Luo, and M. Wang. LINDEN: Linking named entities with knowledge base via semantic knowledge. In *Proceedings of the 21st Conference on World Wide Web*, pages 449–458, 2012.

[129] K. Siorpaes and M. Hepp. Games with a purpose for the semantic web. *Intelligent Systems, IEEE*, 23(3):50 –60, may-june 2008.

[130] Meredith M. Skeels and Jonathan Grudin. When social networks cross boundaries: A case study of workplace use of Facebook and LinkedIn. In *Proceedings of the ACM 2009 international conference on Supporting group work*, GROUP '09, pages 95–104, New York, NY, USA, 2009. ACM.

[131] G. Solskinnsbakk and J. A. Gulla. Semantic annotation from social data. In *Proceedings of the Fourth International Workshop on Social Data on the Web Workshop*, 2011.

[132] Markus Strohmaier, Christian Koerner, and Roman Kern. Why do users tag? detecting users' motivation for tagging in social tagging systems. 2010.

[133] M. Szomszor, H. Alani, I. Cantador, K. O'Hara, and N. Shadbolt. Semantic modelling of user interests based on cross-folksonomy analysis. In *Proceedings of the 7th International Conference on The Semantic Web (ISWC)*, pages 632–648. Springer-Verlag, 2008.

[134] V. Tablan, I. Roberts, H. Cunningham, and K. Bontcheva. Gatecloud.net: a platform for large-scale, open-source text processing on the cloud. *Philosophical Transactions of the Royal Society A*, In Press.

[135] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 1(September 2010):1–41, 2011.

[136] K. Tao, F. Abel, C. Hauff, and G.-J. Houben. What makes a tweet relevant to a topic. In *Proceedings of the #MSM2012 Workshop, CEUR*, volume 838, 2012.

[137] J. Teevan, D. Ramage, and M. R. Morris. #twittersearch: a comparison of microblog search and web search. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 35–44, 2011.

[138] S. Thaler, K. S. E. Simperl, and C. Hofer. A survey on games for knowledge acquisition. Technical Report Tech. Rep. STI TR 2011-05-01, Semantic Technology Institute, 2011.

[139] J. Waitelonis, N. Ludwig, M. Knuth, and H. Sack. WhoKnows? Evaluating linked data heuristics with a quiz that cleans up DBpedia. *Interactive Technology and Smart Education*, 8(4):236–248, 2011.

[140] A. Weichselbraun, S. Gindl, and A. Scharl. Using games with a purpose and bootstrapping to create domain-specific sentiment lexicons. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM)*, pages 1053–1060, 2011.

[141] J. Y. Weng, C. L. Yang, B. N. Chen, Y. K. Wang, and S. D. Lin. IMASS: An Intelligent Microblog Analysis and Summarization System. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, pages 133–138, Portland, Oregon, 2011.

[142] W Wu, B Zhang, and M Ostendorf. Automatic generation of personalized annotation tags for twitter users. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 689–692, 2010.

[143] W. Xin, Z. Jing, J. Jing, H. Yang, S. Palakorn, W. X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E. P. Lim, and X. Li. Topical keyphrase extraction from Twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 379–388, 2011.

[144] S. Yardi and D. Boyd. Tweeting from the town square: Measuring geographic local networks. In *Proceedings of ICWSM*, 2010.

[145] Omar F. Zaidan and Chris Callison-Burch. Crowdsourcing translation: professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229, 2011.

[146] F. Zanzotto, M. Pennaccchiotti, and K. Tsioutsiouliklis. Linguistic Redundancy in Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 659–669, Edinburgh, UK, 2011. Association for Computational Linguistics.

[147] E. Zavitsanos, G. A. Vouros, and G. Paliouras. Classifying users and identifying user interests in folksonomies. In *Proceedings of the 2nd Workshop on Semantic Personalized Information Management: Retrieval and Recommendation*, 2011.

[148] W. X. Zhao, J. Jiang, J. Weng, J. He, E. Lim, H. Yan, and X. Li. Comparing Twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR)*, pages 338–349, 2011.

[149] M. Zhou, S. Bao, X. Wu, and Y. Yu. An unsupervised model for exploring hierarchical semantics from social annotations. In *Proceedings of the 6th International Semantic Web Conference*, ISWC'07, pages 680–693. Springer-Verlag, 2007.