Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web

Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao

Web Information Systems, Delft University of Technology
{f.abel,q.gao,g.j.p.m.houben,k.tao}@tudelft.nl

Abstract. As the most popular microblogging platform, the vast amount of content on Twitter is constantly growing so that the retrieval of relevant information (streams) is becoming more and more difficult every day. Representing the semantics of individual Twitter activities and modeling the interests of Twitter users would allow for personalization and therewith countervail the information overload. Given the variety and recency of topics people discuss on Twitter, semantic user profiles generated from Twitter posts moreover promise to be beneficial for other applications on the Social Web as well. However, automatically inferring the semantic meaning of Twitter posts is a non-trivial problem.

In this paper we investigate semantic user modeling based on Twitter posts. We introduce and analyze methods for linking Twitter posts with related news articles in order to contextualize Twitter activities. We then propose and compare strategies that exploit the semantics extracted from both tweets and related news articles to represent individual Twitter activities in a semantically meaningful way. A large-scale evaluation validates the benefits of our approach and shows that our methods relate tweets to news articles with high precision and coverage, enrich the semantics of tweets clearly and have strong impact on the construction of semantic user profiles for the Social Web.

Keywords: semantic enrichment, twitter, user profile construction, news, linkage.

1 Introduction and Motivation

With the advent of social networking, tagging or microblogging that become tangible in Social Web systems like Facebook, Delicious and Twitter, a new culture of participation penetrates the Web. Today, more than 190 million people are using Twitter and together publish more than 65 million messages (*tweets*) per day[1]. Recent research shows that the exploitation of tweets allows for valuable applications such as earthquake warning systems [1], opinion mining [2] or discovery and ranking of fresh Web sites [3]. These applications mainly analyze and utilize the wisdom of the crowds as source of information rather than relying on individual tweets. Analogously, previous research in the field of microblogging

[1] http://techcrunch.com/2010/06/08/twitter-190-million-users/

G. Antoniou et al. (Eds.): ESWC 2011, Part II, LNCS 6644, pp. 375–389, 2011.

studied information propagation patterns in the global Twitter network [4,5] and exploited network structures to identify influentual users [6,7] as well as malicious users [8,9]. While related work reveals several insights regarding the characteristics of the global Twitter network, there exists little research on understanding the semantics of individual microblogging activities and modeling individual users on Twitter with Semantic Web technologies.

Learning and modeling the semantics of individual Twitter activities is important because the amount of tweets published each day is continuously growing so that users need support to benefit from Twitter information streams. For example, given the huge amount of information streams available on Twitter, user profiling and personalization techniques that support users in selecting streams to follow or particular items to read are becoming crucial [6]. Further, given the variety and recency of topics people discuss on Twitter [5], user profiles that capture the semantics of individual tweets are becoming interesting for other applications on the Social Web as well. In order to provide personalization functionalities in Twitter and moreover enable Social Web applications to consume semantically meaningful representations of the users' Twitter activities, there is thus urgent need to research user modeling strategies that allow for the construction of user profiles with rich semantics.

In this paper we introduce approaches for enriching the semantics of Twitter posts and modeling users based on their microblogging activities. Given the realtime nature and news media characteristics of Twitter [5], we explore possibilities of linking Twitter posts with news articles from the Web. Therefore, we present and evaluate different strategies that link tweets with news articles and contextualize the semantics of tweets with semantics extracted from the corresponding news articles. Based on a large dataset of more than 3 million tweets published by more than 45,000 users, we compare the performance of different strategies for linking tweets with news and analyze their impact on user modeling.

2 Related Work

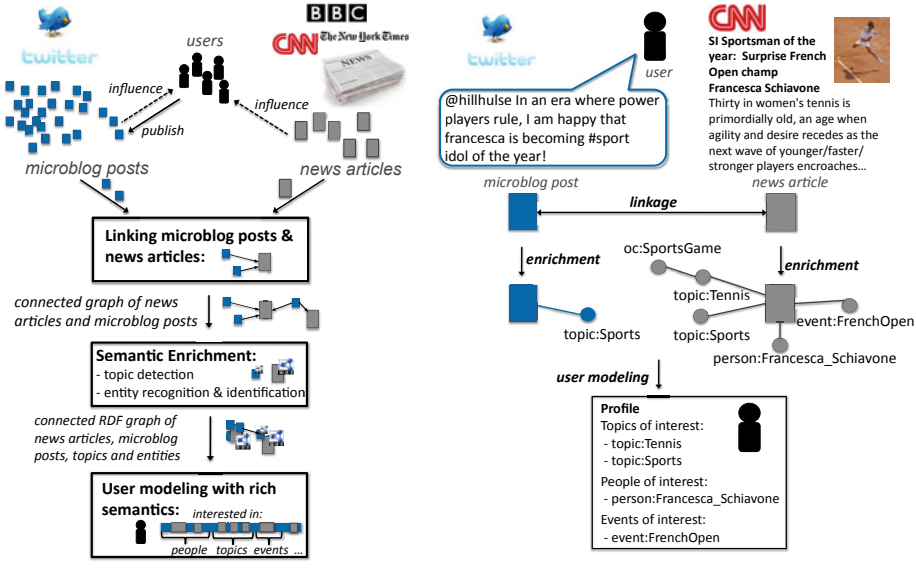
Since Twitter was launched in 2007 research started to investigate the phenomenon of microblogging. Most research on Twitter investigates the network structure and properties of the Twitter network, e.g. [4,5,6,7]. Kwak et al. conducted a temporal analysis of trending topics in Twitter and discovered that over 85% of the tweets posted everyday are related to news [5]. They also show that hashtags are good indicators to detect events and trending topics. Huang et al. analyze the semantics of hashtags in more detail and reveal that tagging in Twitter rather used to join public discussions than organizing content for future retrieval [10]. Laniada and Mika [11] have defined metrics to characterize hashtags with respect to four dimensions: frequency, specificity, consistency, and stability over time. The combination of measures can help assessing hashtags as strong representative identifiers. Miles explored the retrieval of hashtags for recommendation purposes and introduced a method which considers user interests

in a certain topic to find hashtags that are often applied to posts related to this topic [12]. In this paper, we compare hashtag-based methods with methods that extract and analyze the semantics of tweets. While SMOB [13], the semantic microblogging framework, enables users to explicitly attach semantic annotations (URIs) to their short messages by applying MOAT [14] and therewith allows for making the meaning of (hash)tags explicit, our ambition is to infer the semantics of individual Twitter activities automatically.

Research on information retrieval and personalization in Twitter focussed on ranking users and content. For example, Cha et al. [7] present an in-depth comparison of three measures of influence, in-degree, re-tweets, and mentions, to identify and rank influential users. Based on these measures, they also investigate the dynamics of user influence across topics and time. Weng et al. [6] focus on identifying influential users of microblogging services as well. They reveal that the presence of reciprocity can be explained by phenomenon of homophily, i.e. people who are similar are likely to follow each other. Content recommendations in Twitter aim at evaluating the importance of information for a given user and directing the user's attention to certain items. Anlei et al. [3] propose a method to use microblogging streams to detect fresh URLs mentioned in Twitter messages and compute rankings of these URLs. Chen et al. also focus on recommending URLs posted in Twitter messages and propose to structure the problem of content recommendations into three separate dimensions [15]: discovering the source of content, modeling the interests of the users to rank content and exploiting the social network structure to adjust the ranking according to the general popularity of the items. Chen et al. however do not investigate user modeling in detail, but represent users and their tweets by means of a bag of words, from which they remove stop-words. In this paper we go beyond bag-of-word representations and link tweets to news articles from which we extract entities to generate semantically more meaningful user profiles.

Interweaving traditional news media and social media is the goal of research projects such as SYNC3[2], which aims to enrich news events with opinions from the blogosphere. Twitris 2.0 [16] is a Semantic Web platform that connects event-related Twitter messages with other media such as YouTube videos and Google News. Using Twarql [17] for the detection of DBpedia entities and making the semantics of hashtags explicit (via *tagdef*[3]), it captures the semantics of major news events. TwitterStand [18] also analyzes the Twitter network to capture tweets that correspond to late breaking news. Such analyses on certain news events, such as the election in Iran 2009 [2] or the earthquake in Chile 2010 [19], have also been conducted by other related work. However, analyzing the feasibility of linking individual tweets with news articles for enriching and contextualizing the semantics of user activities on Twitter to generate valuable user profiles for the Social Web – which is the main contribution of this paper – has not been researched yet.

[2] `http://www.sync3.eu`
[3] `http://tagdef.com`



(a) Generic solution (b) Example scenario

Fig. 1. Generic solution for semantic enrichment of tweets and user profile construction: (a) generic architecture and (b) example of processing tweets and news articles

3 How to Exploit Twitter for Semantic User Modeling?

The length of Twitter messages is limited to 140 characters which makes it difficult to detect the semantics of these messages. For example, posts such as "Interesting: http://bit.ly/iajV21 #politics" or "@nytimes this makes me scared" are even for humans difficult to understand without knowing the context. However, by following the links one can explore this context and grasp the semantics of the tweets. Many Twitter activities are related to news events. According to Kwak et al., more than 85% of the tweets in the Twitter network are related to news [5]. This observation motivates our idea of linking Twitter activities with news articles to automatically capture and enrich the semantics of microblogging activities. Such relations between tweets and news further allow for capturing user interests regarding trending topics and support applications that require recent user interests like recommender systems for news or other fresh items.

Figure 1(a) visualizes the components of our approach for constructing user profiles with rich semantics based on Twitter posts. We relate Twitter messages with news articles and exploit the content of both tweets and news articles to derive the semantics of the users' microblogging activities. Therefore, we aggregate individual posts of Twitter users as well as news articles published by mainstream media such as CNN, BBC, or New York Times and propose the following components.

Linkage. The challenge of linking tweets and news articles is to identify these articles a certain Twitter message refers to. Sometimes, users explicitly link to the corresponding Web sites, but often there is no hyperlink within a Twitter message which requires more advanced strategies. In Section 4 we introduce and evaluate different strategies that allow for the discovery of relations between tweets and news articles.

Semantic Enrichment. Given the content of tweets and news articles, another challenge is to extract valuable semantics from the textual content. Further, when processing news article Web sites an additional challenge is to extract the main content of the news article. While RSS facilitates aggregation of news articles, the main content of a news article is often not embedded within the RSS feed, but is available via the corresponding HTML-formatted Web site. These Web sites contain supplemental content (*boilerplate*) such as navigation menus, advertisements or comments provided by readers of the article. To extract the main content of news articles we use BoilerPipe [20], a library that applies linguistic rules to separate main content from the boilerplate.

In order to support user modeling and personalization it is important to – given the raw content of tweets and news articles – distill topics and extract entities users are concerned with. We therefore utilize Web services provided by OpenCalais[4], which allow for the extraction of entities such as people, organizations or events and moreover assign unique URIs to known entities and topics.

The connections between the semantically enriched news articles and Twitter posts enable us to construct a rich RDF graph that represents the microblogging activities in a semantically well-defined context.

User Modeling. Based on the RDF graph, which connects Twitter posts, news articles, related entities and topics, we introduce and analyze user modeling strategies that create semantically rich user profiles describing different facets of the users (see Section 5).

Figure 1(b) further illustrates our generic solution by means of an example taken from our dataset: a user is posting a message about the election of the sportsman of the year and states that she supports Francesca Schiavone, an Italian tennis player. The Twitter message itself just mentions the given name *francesca* and indicates with a hashtag (*#sport*) that this post is related to sports. Hence, given just the text from this Twitter message it is not possible to automatically infer that the user is concerned with the tennis player. Given our linkage strategies (see Section 4), one can relate the Twitter message with a corresponding news article published by CNN, which details on the SI sportsman election and Francesca Schiavone in particular. Entity and topic recognition reveal that the article is about tennis (*topic:Tennis*) and Schiavone's (*person:Francesca_Schiavone*) success at French Open (*event:FrenchOpen*) and therewith enrich the semantics which can be extracted from the Twitter message itself (*topic:Sports*).

[4] `http://www.opencalais.com`

4 Analyzing Linkage between Tweets and News for Semantic Enrichment

The key idea of our approach to enrich the semantics of Twitter messages is based on relating individual tweets to news articles so that semantics extracted from news articles can be applied to clarify the meaning of tweets. In this section we introduce different strategies for linking Twitter posts with news articles that provide details on these posts. To evaluate the impact of these strategies on the semantic enrichment of Twitter posts we conduct an analysis on a large dataset gathered from Twitter and major news publishing Web sites.

4.1 Strategies for Discovering Tweet-News Relations

The strategies, which we propose to find correlations between Twitter posts and external news resources, can be divided into URL-based strategies, which exploit interaction patterns and hyperlinks mentioned in tweets, and content-based strategies, which exploit the content of tweets and news articles. In the following definitions, T denotes the set of all tweets available in our dataset while N refers to the set of news articles.

URL-based Strategies. URLs (mostly short URLs shortened by services such as *bit.ly*) that are contained in tweets can be considered as indicators for news-related tweets. In particular, if a tweet contains a URL that points to an external news resource, there is a very high possibility that this tweet is closely related to the linked resource. Based on this principle we defined two URL-based strategies.

Definition 1 (Strict URL-based strategy). *If a Twitter post $t \in T$ contains at least one URL that is from certain mainstream news publishers and links to a news article $n \in N$, then we consider t and n as related: $(t, n) \in R_s$, where $R_s \subseteq T \times N$.*

For this strategy, we select BBC, CNN and the New York Times as the set of mainstream news publishers and apply URL-patterns to discover the corresponding tweets that point to these Web sites. A potential drawback of the strict URL-based strategy is that it will miss relevant relations for Twitter messages that contains no URL. For example, if a user replies to Twitter message that is according to the strict URL-based strategy related to a news article n then this reply message might be related to n as well. Based on this idea, we define a second URL-based strategy that is more flexible than the first one.

Definition 2 (Lenient URL-based strategy). *If a tweet $t_r \in T$ is a reply or re-tweet from another tweet $t \in T$, which contains at least one URL that is linked to a news article $n \in N$ authored by certain mainstream news publishers, then we consider both t_r and t as being related to n: $(t_r, n) \in R_l, (t, n) \in R_l$, where $R_l \subseteq T \times N$.*

Hence, the lenient URL-based strategy extends the strict strategy with tweets that were published as part of an interaction with a tweet that is according to the strict strategy news-related so that $R_s \subseteq R_l$.

Content based Strategies. As tweets do not necessarily contain a URL, we propse another set of strategies that exploit the content of tweets and news articles to connect tweets with news. For example, the Twitter post about Francesca Schiavone in Figure 1(b) should be linked to the corresponding news article even though the tweet does not have a URL directly pointing to the article. We thus propose three further strategies that analyze the content of Twitter posts to allow for linkage between Twitter activities and news articles.

Definition 3 (Bag-of-Words Strategy). *Formally, a Twitter post $t_j \in T$ can be represented by a vector $\boldsymbol{t} = (\alpha_1, \alpha_2..\alpha_m)$ where α_i is the frequency of a word i in t and m denotes the total number of words in t. Each news article $n \in N$ is also represented as a vector $\boldsymbol{n} = (\beta_1, \beta_2..\beta_k)$ where β_i is the frequency of a word i in the title of the news article n and k denotes the total number of words in n.*

The bag-of-word strategy relates a tweet t with the news article n, for which the $TF \times IDF$ score is maximized: $(t, n) \in R_b$, where $R_b \subseteq T \times N$.

The bag-of-words strategy thus compares a tweet t with every news article in N and chooses the most similar ones to build a relation between t and the corresponding article n. $TF \times IDF$ is applied to measure the similarity. Given a Twitter post t and a news article n, the term frequency TF_i of a term i (with $\alpha_i > 0$ in the vector representation of t) is β_i, i.e. the number of occurrences of the word i in n. And IDF_i, the inverse document frequency, is $IDF_i = 1 + log(\frac{|N|}{|\{n \in N : \beta_i > 0\}|+1})$, where $|\{n \in N : \beta_i > 0\}|$ is the number of news articles, in which the term i appears. Given TF and IDF, the similarity between t and n is calculated as follows.

$$sim(t, n) = \sum_{i=1}^{m} TF_i \cdot IDF_i \tag{1}$$

Given a ranking according to the above similarity measure, we select top ranked tweet-news pairs (t, n) as candidates for constructing a valid relation. Following the realtime nature of Twitter, we also add a temporal constraint to filter out these candidates, for which the publishing date of the Twitter message and news article differs more than two days.

The bag-of-words strategy treats all words in a Twitter post as equally important. However, in Twitter, hashtags can be considered as special words that are important features to characterize a tweet [11]. For news articles, some keywords such as person names, locations, topics, etc. are also good descriptors to characterize a news article. Conveying these observations, we introduce hashtag-based and entity-based strategies for discovering relations between tweets and news articles. These strategies follow the idea of the bag-of-words strategy (see Definition 3) and differ in the way of representing news articles and tweets.

Definition 4 (Hashtag-based strategy). *The hashtag-based strategy represents a Twitter post $t \in T$ via its hashtags: $\boldsymbol{h} = (\alpha_1, \alpha_2..\alpha_m)$, where α_i is the number of occurrences of a hashtag i in t and m denotes the total number of hashtags in t.*

The hashtag-based strategy relates a tweet t (represented via its hashtags) with the news article n, for which the $TF \times IDF$ score is maximized: $(t, n) \in R_h$, *where* $R_h \subseteq T \times N$.

While the hashtag-based strategy thus varies the style of representing Twitter messages, the entity-based strategy introduces a new approach for representing news articles.

Definition 5 (Entity-based strategy). *Twitter posts* $t \in T$ *are represented by a vector* $\boldsymbol{t} = (\alpha_1, \alpha_2..\alpha_m)$ *where* α_i *is the frequency of a word i in t and m denotes the total number of words in t. Each news article* $n \in N$ *is represented by means of a vector* $\boldsymbol{n} = (\beta_1, \beta_2..\beta_k)$, *where* β_i *is the frequency of an entity within the news article, i is the label of the entity and k denotes the total number of distinct entities in the news article n.*

The entity-based strategy relates the Twitter post t (represented via bag-of-words) with the news article n (represented via the labels of entities mentioned in n), for which the $TF \times IDF$ *score is maximized:* $(t, n) \in R_e$, *where* $R_e \subseteq T \times N$.

Entities are extracted by exploiting OpenCalais as described in Section 3. For the hashtag- and entity-based strategies, we thus use Equation 1 to generate a set of candidates of related tweet-news pairs and then filter out these pairs, which do not fulfill the temporal constraint that prescribes that the tweet and news article should be published within a time span of two days. Such temporal constraints may reduce the recall but have a positive effect on the precision as we will see in our analysis below.

4.2 Analysis and Evaluation

To analyze the impact of the strategies on semantic enrichment of Twitter posts, we evaluate the performance of the strategies with respect to coverage and precision based on a large data corpus which we crawled from Twitter and three major news media sites: BBC, CNN and New York Times.

Data Collection and Characteristics. Over a period of three weeks we crawled Twitter information streams via the Twitter streaming API. We started from a seed set of 56 Twitter accounts (U_n), which are maintained by people associated with one of the three mainstream news publishers, and gradually extended this so that we finally observed the Twitter activities of 48,927 extra users (U_u), who are not explicitly associated with BBC, CNN or the New York Times. The extension was done in a snowball manner: we added users to U_u, who interacted with another user $u \in U_n \cup U_u$. The 56 Twitter accounts closely related to mainstream news media enable publishers of news articles to discuss their articles and news events with the Twitter audience. For the 48,927 *casual users* we crawled all Twitter activities independently whether the activity was part of an interaction with the Twitter accounts of the mainstream news media or not. In total we thereby obtained more than 3.3 million tweets.

Figure 2 shows the number of tweets per user and depicts how often these users interacted with a Twitter account associated with mainstream media. The

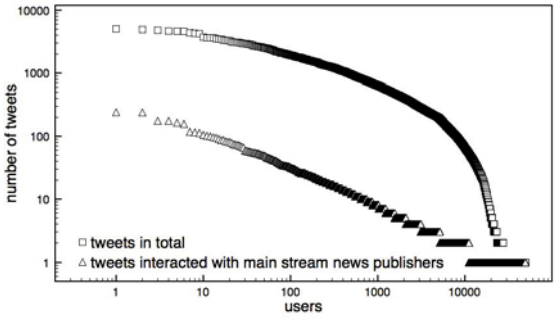


Fig. 2. Number of tweets per user $u \in U_u$ as well as the number of interactions (re-tweeting or reply activities) with Twitter accounts maintained by mainstream news media

distribution of the number of tweets per user shows a power-law-like distribution. For many users we recorded less than 10 Twitter activities within the three week observation period. Less than 500 users were highly active and published more than 1000 tweets. Further, the majority (more than 75%) of users interacted only once with a news-related user $u \in U_n$. We observed only nine users, who re-tweeted or replied to messages from news-related users more than 100 times and identified one of these users as spam user, who just joined the discussion to promote Web sites.

To connect tweets with news articles, we further crawled traditional news media. Each of the three mainstream news publishers (BBC, CNN, and New York Times) also provides a variety of news channels via their Web site. These news channels correspond to different news categories such as politics, sports, or culture and are made available via RSS feed. We constantly monitored 63 different RSS feeds from the corresponding news publishers and crawled the main content as well as supplemental metadata (title, author, publishing data, etc.) of more than 44,000 news articles.

Experimental Results. In order to evaluate the accuracy of the different strategies for relating tweets with news articles, we randomly selected tweet-news pairs that were correlated by a given strategy and judged the relatedness of the news article and the tweet message on a scale between 1 ("not related") and 4 ("perfect match"), where 2 means "not closely related" and 3 denotes "related" tweet-news pairs. For example, given a Twitter message about Francesca Schiavone's victory at French Open 2010, we considered news articles that report about this victory as "perfect match" while news articles about Francesca Schiavone, for which this victory is not the main topic but just mentioned as background information were considered as "related". In total, we (the authors of this paper) judged 1427 tweet-news pairs where each of the strategies depicted in Figure 3 was judged at least 200 times. For 85 pairs (5.96%) we were not able to decide whether the corresponding Twitter posts and the news article are related. We considered these pairs as "not related" and tweet-news relations, which were rated at least with 3 as truly related.

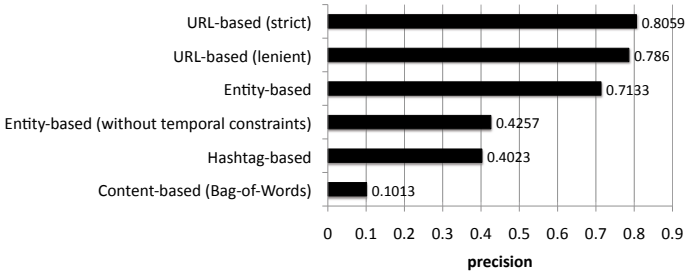


Fig. 3. Precision of different strategies for relating Twitter messages with news articles. (considered to refer to accurate).

Given this ground truth of correct tweet-news relations, we compare the precision of the different strategies, i.e. the fraction of correctly generated tweet-news relations. Figure 3 plots the results and shows that the URL-based strategies perform best with a precision of 80.59% (strict) and 78.8% (lenient) respectively. The naive content-based strategy, which utilizes the entire Twitter message (excluding stop-words) as search query and applies TFxIDF to rank the news articles, performs worst and is clearly outperformed by all other strategies. It is interesting to see that the entity-based strategy, which considers the publishing date of the Twitter message and news article, is nearly as good as the lenient URL-based strategy and clearly outperforms the hashtag-based strategy, which uses the temporal constraints as well. Even without considering temporal constraints, the entity-based strategy results in higher accuracy than the hashtag-based strategy. We conclude that the constellation/set of entities mentioned in a news article and Twitter message correspondingly, i.e. the number of shared entities, is a good indicator of relating tweets and news articles.

Figure 4 shows the coverage of the strategies, i.e. the number of tweets per user, for which the corresponding strategy found an appropriate news article. The URL-based strategies, which achieve the highest accuracy, are very restrictive: for less than 1000 users the number of tweets that are connected to news articles is higher than 10. The coverage of the lenient URL-based strategy is clearly higher than for the strict one, which can be explained by the number of interactions with Twitter accounts from mainstream news media (see Figure 2). The hashtag-based and entity-based strategies even allow for a far more higher number of tweet-news pairs. However, the hashtag-based strategy fails to relate tweets for more than 79% of the users, because most of these people do not make use of hashtags. By contrast, the entity-based strategy is applicable for the great majority of people and, given that it showed an accuracy of more than 70% can be considered as the most successful strategy.

Combining all strategies results in the highest coverage: for more than 20% of the users, the number of tweet-news relations is higher than 10. In the next section we will show that given these tweet-news relations we can create rich profiles that go beyond the variety of profiles, which are just constructed based on the tweets of the users.

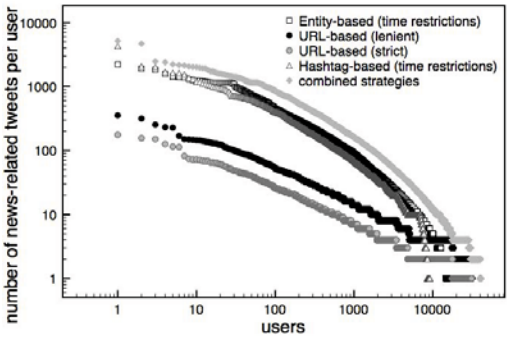


Fig. 4. Number of tweets per user, which are according to the different strategies related to news articles

5 Analyzing User Profile Construction Based on Semantic Enrichment

Based on the linkage of Twitter activities with news articles, we can exploit the semantics embodied in the news articles to create and enrich user profiles. In this section, we first present approaches for user modeling based on Twitter activities and then analyze the impact of exploiting related news articles for user profile construction in Twitter.

5.1 User Modeling Strategies

In this study we focus on two types of profiles: entity-based and topic-based profiles. An entity-based profile models a user's interests into a given set of entities such as persons, organizations, or events and can be defined as follows.

Definition 6 (Entity-based profile). *The* entity-based profile *of a user* $u \in U$ *is a set of weighted entities where the weight of an entity* $e \in E$ *is computed by a certain strategy* w *with respect to the given user* u*.*

$$P(u) = \{(e, w(u, e)) | e \in E, u \in U\} \tag{2}$$

$w(u, e)$ *is the weight that is associated with an entity* e *for a given user* u*.* E *and* U *denote the set of entities and users respectively.*

In Twitter, a naive strategy for computing a weight $w(u, e)$ is to count the number of u's tweets that refer to the given entity e. $|P(u)|$ depicts the number of distinct entities that appear in a profile $P(u)$. While entity-based profiles represent a user in a detailed and fine-grained fashion, topic-based profiles describe a user's interests into topics such as sports, politics or technology that can be specified analogously (see Definition 7).

Definition 7 (Topic-based profile). *The* topic-based profile $P_T(u)$ *of a user* $u \in U$ *is the restriction of an entity-based profile* $P(u)$ *to a set of topics* $T \subseteq E$*.*

From a technical point of view, both types of profiles specify the interest of a user into a certain URI, which represents an entity or topic respectively. Given the URI-based representation, the entity- and topic-based profiles become part of the Web of Linked Data and can therewith not only be applied for personalization purposes in Twitter (e.g., recommendations of tweet messages or information streams to follow) but in in other systems as well. For the construction of entity- and topic-based profiles we compare the following two strategies.

Tweet-based. The tweet-based baseline strategy constructs entity- and topic-based user profiles by considering only the Twitter messages posted by a user, i.e. the first step of our user modeling approach depicted in Figure 1(a) is omitted so that tweets are not linked to news articles. Entities and topics are directly extracted from tweets using OpenCalais. The weight of an entity corresponds to the number of tweets, from which an entity was successfully extracted, and the weight of a topic corresponds to the number of tweets, which were categorized with the given topic.

News-based. The news-based user modeling strategy applies the full pipeline of our architecture for constructing the user profiles (see Figure 1(a)). Twitter messages are linked to news articles by combining the URL-based and entity-based (with temporal restrictions) strategies introduced in Section 4 and entities and topics are extracted from the news articles, which have been linked with the Twitter activities of the given user. The weights correspond again to the number of Twitter activities which relate to an entity and topic respectively.

Our hypothesis is that the news-based user modeling strategy, which benefits from the linkage of Twitter messages with news articles, creates more valuable profiles than the tweet-based strategy.

5.2 Analysis and Evaluation

To validate our hypothesis we randomly selected 1000 users (from U_u) and applied both strategies to create semantic user profiles from their Twitter activities. Figure 5 compares the number of distinct entities and topics available in the corresponding profiles ($|P(u)|$). Even though the number of Twitter activities, which can be linked to news articles, is smaller than the total number of Twitter activities of a user (cf. Fig. 2, Fig. 4), the number of entities and topics available in the profiles generated via the news-based strategy is higher than for the tweet-based approach. Regarding the entity-based profiles this difference is higher than for the topic-based profiles, because each Twitter message and news article is usually categorized with one topic at most whereas for the number of entities there is no such limit. News articles provide much more background information (a higher number of entities) than Twitter messages and thus allow for the construction of more detailed entity-based user profiles.

Further, the variety of the entity-based profiles generated via the news-based strategy is much higher than for the tweet-based strategy as depicted in Figure 5(c). For the tweet-based strategy, more than 50% of the profiles contain

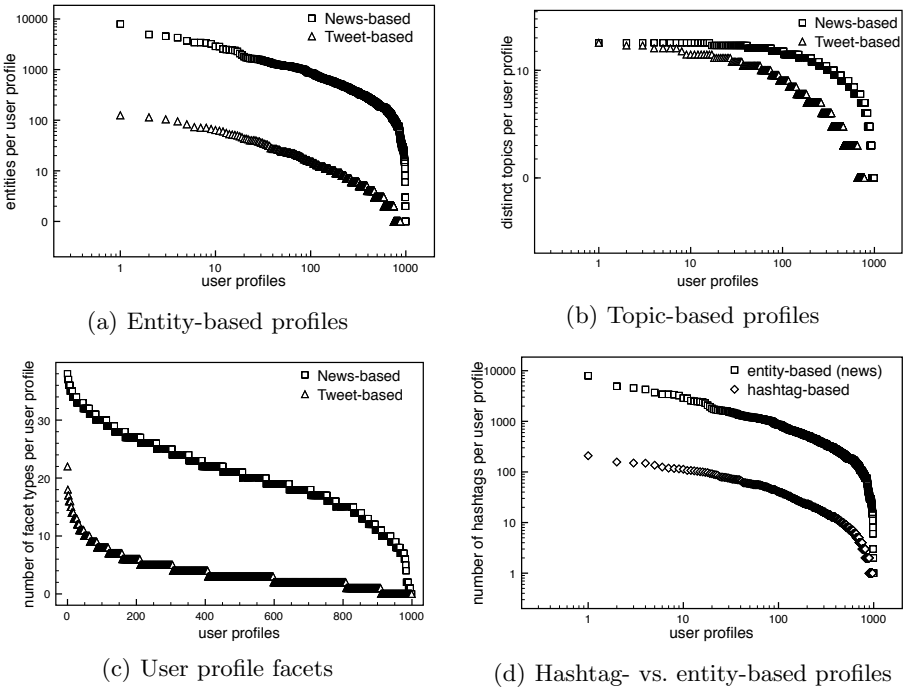


Fig. 5. Comparison between tweet-based and news-based user modeling strategies: (a) for creating entity-based profiles and (b) topic-based profiles, (c) with respect to the variety of facet types available in the user profiles (example facet types: person, event, location, product). Further, (d) hashtag-based vs. entity-based profiles: number of distinct hash tags and entities per profile.

just less than four types of entities (mostly persons and organizations) while for the news-based strategy more than 50% of the profiles reveal interests in more than 20 types of entities. For example, they show that users are – in addition to persons or organizations – also concerned with certain events or products. The news-based strategy, i.e. the complete user construction pipeline proposed in Figure 1, thus allows for the construction of profiles that cover different facets of interests which increases the number of applications that can be built on top of our user modeling approaches (e.g., product recommendations).

Related research stresses the role of hashtags for being valuable descriptors [11,12,10]. However, a comparison between hashtag-based profiles and entity-based profiles created via the news-based strategy shows that for user modeling on Twitter, hashtags seem to be a less valuable source of information. Figure 5(d) reveals that the number of distinct hashtags available in the corresponding user profiles is much smaller than the number of distinct entities that are discovered with our strategy, which relates Twitter messages with news articles. Given that each named entity as well as each topic of an entity- and topic-based user profile has a URI, the semantic expressiveness of profiles generated with the news-based user modeling strategy is much higher than for the hashtag-based profiles.

6 Conclusions and Future Work

In this article, we introduced and analyzed strategies that enrich the semantics of microblogging activities for creating semantically rich user profiles on the Social Web. We present different strategies that connect Twitter messages with related news articles and exploit semantics extracted from news articles to deduce and contextualize the semantic meaning of individual Twitter posts. Our evaluation on a large Twitter dataset (more than 3 million tweets posted by more than 45,000 users) showed that, given the name of entities mentioned in a news article (such as persons or organizations) as well as the temporal context of the article, we can relate tweets and news articles with high precision (more than 70%) and high coverage (approx. 15% of the tweets can be linked to news articles). Our analysis further revealed that the exploitation of tweet-news relation has significant impact on user modeling and allows for the construction of more meaningful profiles (more profile facets and more detailed knowledge regarding user interests/concerns) than user modeling based on tweets only.

Semantic enrichment of Twitter user activities based on semantics extracted from news articles thus leads to meaningful representations of Twitter activities, ready for being applied in Twitter and other Social Web systems. In our ongoing research, we would deepen the investigation of how the profiles constructed by this type of user modeling strategies impact personalization on the Social Web[5]. Given the variety and recency of the constructed profiles, there are different applications worthwhile to explore such as Twitter stream and message recommendations, product recommendations or recommending news.

Acknowledgements. This work is partially sponsored by the EU FP7 projects ImREAL (http://imreal-project.eu) and GRAPPLE (http://grapple-project.org).

References

1. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proc. of 19th Int. Conf. on World Wide Web, pp. 851–860. ACM, New York (2010)
2. Gaffney, D.: #iranElection: quantifying online activism. In: Proc. of the WebSci10: Extending the Frontiers of Society On-Line (2010)
3. Dong, A., Zhang, R., Kolari, P., Bai, J., Diaz, F., Chang, Y., Zheng, Z., Zha, H.: Time is of the essence: improving recency ranking using Twitter data. In: Proc. of 19th Int. Conf. on World Wide Web, pp. 331–340. ACM, New York (2010)
4. Lerman, K., Ghosh, R.: Information contagion: an empirical study of spread of news on Digg and Twitter social networks. In: Cohen, W.W., Gosling, S. (eds.) Proc. of 4th Int. Conf. on Weblogs and Social Media. AAAI Press, Menlo Park (2010)
5. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proc. of the 19th Int. Conf. on World Wide Web, pp. 591–600. ACM, New York (2010)

[5] Code and further results: `http://wis.ewi.tudelft.nl/umap2011/`

6. Weng, J., Lim, E.P., Jiang, J., He, Q.: TwitterRank: finding topic-sensitive influential Twitterers. In: Davison, B.D., Suel, T., Craswell, N., Liu, B. (eds.) Proc. of 3rd ACM Int. Conf. on Web Search and Data Mining, pp. 261–270. ACM, New York (2010)
7. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P.K.: Measuring user influence in twitter: The million follower fallacy. In: Cohen, W.W., Gosling, S. (eds.) Proc. of 4th Int. Conf. on Weblogs and Social Media. AAAI Press, Menlo Park (2010)
8. Lee, K., Caverlee, J., Webb, S.: The social honeypot project: protecting online communities from spammers. In: Proc. of 19th Int. Conf. on World Wide Web, pp. 1139–1140. ACM, New York (2010)
9. Lee, K., Caverlee, J., Webb, S.: Uncovering social spammers: social honeypots + machine learning. In: Proc. of 33rd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 435–442. ACM, New York (2010)
10. Huang, J., Thornton, K.M., Efthimiadis, E.N.: Conversational tagging in twitter. In: Proc. of 21st Conf. on Hypertext and Hypermedia, pp. 173–178. ACM, New York (2010)
11. Laniado, D., Mika, P.: Making sense of twitter. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 470–485. Springer, Heidelberg (2010)
12. Efron, M.: Hashtag retrieval in a microblogging environment. In: Proc. of 33rd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 787–788. ACM, New York (2010)
13. Passant, A., Hastrup, T., Bojars, U., Breslin, J.: Microblogging: A Semantic Web and Distributed Approach. In: Bizer, C., Auer, S., Grimnes, G.A., Heath, T. (eds.) Proc. of 4th Workshop Scripting For the Semantic Web (SFSW 2008) co-located with ESWC 2008, vol. 368 (2008), CEUR-WS.org
14. Passant, A., Laublet, P.: Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data. In: Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW 2008), Beijing, China (2008)
15. Chen, J., Nairn, R., Nelson, L., Bernstein, M., Chi, E.: Short and tweet: experiments on recommending content from information streams. In: Proc. of 28th Int. Conf. on Human Factors in Computing Systems, pp. 1185–1194. ACM, New York (2010)
16. Jadhav, A., Purohit, H., Kapanipathi, P., Ananthram, P., Ranabahu, A., Nguyen, V., Mendes, P.N., Smith, A.G., Cooney, M., Sheth, A.: Twitris 2.0: Semantically empowered system for understanding perceptions from social data. In: Proc. of the Int. Semantic Web Challenge (2010)
17. Mendes, P.N., Passant, A., Kapanipathi, P.: Twarql: tapping into the wisdom of the crowd. In: Proc. of the 6th International Conference on Semantic Systems, pp. 45:1–45:3. ACM, New York (2010)
18. Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., Sperling, J.: Twitterstand: news in tweets. In: Proc. of 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 42–51. ACM, New York (2009)
19. Mendoza, M., Poblete, B., Castillo, C.: Twitter Under Crisis: Can we trust what we RT? In: Proc. of 1st Workshop on Social Media Analytics (SOMA 2010). ACM Press, New York (2010)
20. Kohlschütter, C., Fankhauser, P., Nejdl, W.: Boilerplate detection using shallow text features. In: Proc. of 3rd ACM Int. Conf. on Web Search and Data Mining, pp. 441–450. ACM, New York (2010)