Learning Semantic Relationships between Entities in Twitter

Ilknur Celik, Fabian Abel, and Geert-Jan Houben

Web Information Systems, Delft University of Technology
Mekelweg 4, 2628 Delft, The Netherlands
{i.celik,f.abel,g.j.p.m.houben}@tudelft.nl

Abstract. In this paper, we investigate whether semantic relationships between entities can be learnt from analyzing microblog posts published on Twitter. We identify semantic links between persons, products, events and other entities. We develop a relation discovery framework that allows for the detection of typed relations that moreover may have temporal dynamics. Based on a large Twitter dataset, we evaluate different strategies and show that co-occurrence based strategies allow for high precision and perform particularly well for relations between persons and events achieving precisions of more than 80%. We further analyze the performance in learning relationships that are valid only for a certain time period and reveal that for those types of relationships Twitter is a suitable source as it allows for discovering trending topics with higher accuracy and with lower delay in time than traditional news media.

Keywords: semantic enrichment, relation learning, twitter, social web.

1 Introduction

Twitter has become an increasingly popular social network tool, not only for socializing with friends [6], but also for gathering information about what is going on in the world in whatever subject one might be interested in [8,17,7]. People typically post short status messages, interesting information they read from daily news, updates on current hot topics, their views and opinions on a certain subject matter, developments in their profession, materials about their interests, et cetera [16,12,23]. It is a growing real-time information network with several hundreds of millions of users and millions of tweets per day[1]. It is estimated that highly active users regularly receive around one thousand tweets every day [4]. This information overload may cause users to get lost in the information network and become de-motivated. Finding your way around Twitter is indeed not very straightforward due to the lack of a user-friendly browsing option that would allow for a more faceted browsing experience than the existing chronologically-ordered clutter option [4,19]. Accessing required or interested fresh content easily is vital in today's information age. Hence, there is a need for guidance from users'

---

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

point of view that would assist them in finding the related information about a subject they have just read, or where they can browse the information sources by related people, countries, cities, events, selected categories and so on.

Today, browsing and exploring micro-blogging content along certain topics is not easily supported by Twitter and other micro-blogging services. On the other hand, news websites such as Google News[2] provide faceted search interfaces to facilitate exploration of news. Lately, services such as the European Media Monitor[3] start to exploit entities (e.g. persons, locations) for faceted search as well as relations between entities to provide recommendations. For example, if users click on a person they retrieve news related to this person and can see a list of related people. Twitter does not provide such advanced browsing functionality and the feasibility of identifying entities and relationships between entities in Twitter has not been studied yet.

Our research closes this gap and focuses on finding relationships in and between tweets based on entities in order to provide a medium where users can easily navigate through the information network by facets or easily see and access relevant content for what they are interested. For this reason, we need to understand the semantics of tweets in order to infer recommendations and support user navigation. However, Twitter posts are limited to 140 characters only. Identifying entities, and moreover relating entities, via Twitter messages is thus a non-trivial problem. In order to facilitate the identification of entities within tweets, we enrich Twitter messages by mapping them to related news articles [2]. We then apply different strategies to learn relationships between entities within tweets as well as news. Our main focus is directed at determining whether relations can be discovered from Twitter messages or whether some enrichment is needed, rather than finding the optimal strategy for relation learning. Furthermore, we analyze the validity of the discovered relations over time in order to distinguish which entities are related at a given time, and thus provide the most interesting and relevant content at the appropriate time. Our work, hence, involves engineering of a relation discovery framework which can be utilized for building useful applications such as browsing or recommendation support. The main contributions of our work can be summarized as follows:

– We implement a framework for discovering relationships between entities based on Twitter. Our framework extracts typed entities from enriched tweets/news and provides strategies for detecting semantic (trending) relationships between entities. Learnt relations provide a basis for improving browsing and content exploration as well as content recommendations in Twitter.
– Given a large dataset of more than 10 million tweets and 70,000 news articles[4], we analyze and evaluate the framework and the different strategies for detecting relationships. In particular, we:

---

[2] http://news.google.com/

[3] http://emm.newsexplorer.eu/

[4] We make our datasets publicly available via our supporting website [1]

- investigate the precision and recall of the relation detection strategies,
- analyze how the strategies perform for each type of relationships and
- evaluate the quality and speed for discovering trending relationships that possibly have a limited temporal validity.

The paper is structured as follows. In the next section we will discuss related research. In Section 3 we will explain the design dimensions of our Twitter-based relation discovery framework and discuss different applications of the framework (e.g. browsing support). A detailed analysis and evaluation of the framework is presented in Section 4, before we discuss our main findings in Section 5 and conclude in Section 6.

2   Related Work

Twitter is a new social network phenomenon that is attracting interest from different types of people all around the world for different purposes, e.g. politicians, celebrities, journalists, scholars and the like. Over the last few years, Twitter has shown an exponential growth and became the most popular micro-blogging site. In line with this, there has been a growing interest in analyzing Twitter and it effects from a variety of research communities.

A considerable amount of research on micro-blogging and especially on Twitter focused on examining the user behavior and motivation for using Twitter [13,6], as well as user intentions [9]. Some questioned the usability of Twitter especially for novice users [19], studied capturing comprehensive locality information to be used for viewing tweets in the current local position of the mobile device [14], researched discovering and ranking fresh web sites using tweets [8], investigated the social aspects of group polarization over time in Twitter [22], while others exploited Twitter adoption and use in mass convergence and emergency events [12]. Research on temporal dynamics and information propagation on Twitter mostly focused on hashtags or observations of other sets of words in the tweets [16,17,11,20,3]. We scrutinize validity of relations between entities for given time periods and perform a temporal analysis of these relations.

Chen et al. [7] studied content recommendation in Twitter and found out that both topic and social voting (or relevance) are important considerations. They also observed that URLs extracted from the user's close social group is more successful than the most popular ones. Bernstein et al. developed *Eddi*, a prototype of an interactive topic-based browsing interface for Twitter, after observing how the users manage the incoming flood of updates in Twitter [4]. Their prototype interface groups the tweets of a user into topics mentioned explicitly or implicitly, so that users can browse for tweets of interest by topic, filtering out undesired topics. In summary, they simply categorize the tweets in the feed of the user without using any semantics or natural language processing. This approach, however, does not find the relations between the topics or perform any recommendation of related topics. While this provides a means for browsing through a user's own feed by topics, our ambition is to infer relations between
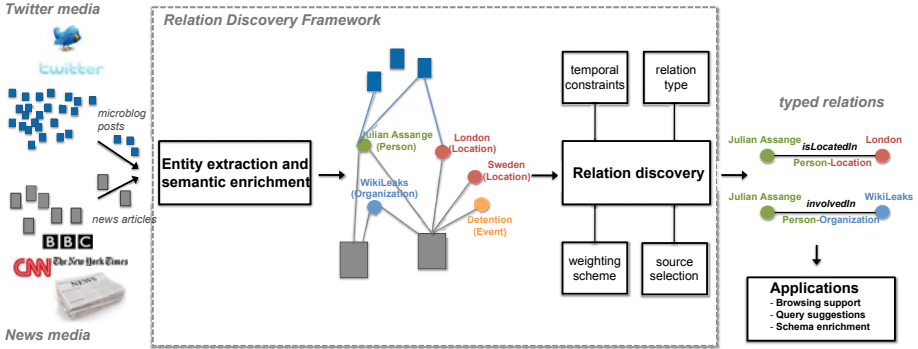
Fig. 1. Conceptual architecture of the Twitter-based relation discovery framework

entities in all tweets in order to present a list that contains the related entities of the tweet of interest even outside of the user's feed.

Tweets are limited to 140 characters and are therewith too short to extract meaningful semantics from them on their own. Furthermore, due to this limited space, users tend to use abbreviations and short-form for words in order to save space. This can only make things harder for inferring semantics from tweets. Rowe et al. [21] mapped tweets to conference talks and exploited metadata of the corresponding research papers to enrich the semantics of tweets to better understand the semantics of the tweets published in conferences. Our approach is somewhat similar to this, except we try to enrich the tweets not in a restricted domain like scientific conference, but in general. For this reason we try to map the Twitter posts we crawled with news articles on the Web over the same time period, and learn relations between them. Marinho et al. [18] proposed a method for collabulary learning which takes a folksonomy and domain-expert ontology as input and performs semantic mapping to generate an enriched folksonomy. An algorithm based on frequent itemsets is then applied to learn an ontology over this enriched folksonomy. A similar approach is presented by Hotho et al. [10] who exploit frequent itemsets to learn association rules from tagging activities. We primarily exploit co-occurrence frequencies of entity pairs and investigate whether Twitter messages in combination with news articles can be exploited to learn relations between these entities. Learning such relationships and structures in Twitter is important to countervail information overload in Twitter and support browsing through Twitter information streams.

3 Framework for Learning Relations between Entities in Twitter

Figure 1 visualizes our approach to discovering relationships between entities. Given posts available in Twitter or the news media, our framework processes these posts and performs two main steps: (1) entity extraction and semantic enrichment and (2) relation discovery.

The first step is described in [2] and allows for the detection of entities and further semantics such as the entity type or the topic a post refers to. This process results in a graph, which connects the semantically enriched resources (tweets and news articles) with entities that are mentioned in the corresponding resources. This graph can be represented via a triple $\mathbb{G} := (R, E, Y)$, where R and E are the set of resources and entities respectively and $Y \subseteq E \times R$ is a relation between entities and resources, i.e. $(e, r) \in Y$ if entity e is referenced from resource r.

In the second step, the relation discovery strategies exploit this graph to detect pairs of entities that have a certain type of relationship (*relation type*) in a specific period of time (*temporal constraints*). Such relationships can be defined as tuples:

Definition 1 (Relationship). *Given two entities e_1 and e_2, a relationship between these entities is described via a tuple $rel(e_1, e_2, type, t_{start}, t_{end}, w)$, where* type *labels the relationship, t_{start} and t_{end} specify the temporal validity of the relationship and $w \in [0..1]$ is a weighting score that allows for specifying the strength of the relationship.*

The higher the weighting score w the stronger the relationship between e_1 and e_2. If two entities are not related then the weight is 0. Given the type of relationships that should be learnt and (possibly) temporal constraints (t_{start} and t_{end}) that prescribe for what time period the strategy should seek for such relationships, there exist two main design dimensions that influence the relation discovery (see Fig. 1).

Source selection. The source selection strategy decides which parts of the available data sources will be exploited when deducing relationships between entities, i.e. it specifies the sub-graph spanned by $\mathbb{G}$, which connects entities and posts (see Fig. 1), that will be exploited to discover relationships. In this paper, we differentiate between three different sources: (i) Twitter messages ($\mathbb{G}_T$), (ii) news articles ($\mathbb{G}_N$) or (iii) a combination of both media ($\mathbb{G}_C$).

Weighting scheme. A core challenge of the relation discovery is to compute the weight w, which expresses the strength of a relationship (see Definition 1). Those pairs of entities that are—according to the given type of the relationship—strongly related should be weighted high while for rather unrelated entities the weight should be low. In this paper, we utilize the co-occurrence frequency of two entities as weighting scheme. Hence, given a certain data source, we count the number of resources (tweets and/or news articles), in which both entities e_1 and e_2 are mentioned: $w_{co}(e_1, e_2) = |r \in R|(e_1, r) \in Y \wedge (e_2, r) \in Y|$.

As depicted in Fig. 1, our Twitter-based relation discovery framework outputs semantic relationships between entities that support various applications. For example, relationships between entities can, in combination with the enriched data $\mathbb{G}$, be applied to enable browsing support by linking to related tweets. Furthermore, when people search in Twitter, our framework can identify entities

mentioned in a keyword query and exploit relationships to other entities to provide query suggestions to the user. In addition to these applications, our relation discovery framework can assist in building ontologies on the fly. As our framework provides functionality to map entities to concepts of existing ontologies such as DBpedia [5], learnt relationships can also be applied to complement and enrich existing knowledge bases and schemata.

We implemented the relation discovery framework as a Java library using the Twitter streaming API[5] to monitor tweets, boilerpipe[6] [15] for extracting the main content of news articles from websites such as BBC or CNN and OpenCalais[7] for extracting entities from tweets and news articles respectively.

4   Evaluation of Strategies for Learning Relationships

Given the relation discovery framework introduced in the previous section, we evaluate the performance of the different strategies and investigate the following research questions.

1. Which media source (Twitter, News or a combination of both) allows for the highest precision and recall in detecting relationships between entities?
2. Does the accuracy depend on the type of entities which are involved in a relation that should be discovered, and, if so, for what type of relations is the accuracy high/low?
3. How do the strategies perform for discovering relationships which have temporal constraints, and how fast can the strategies detect (trending) relationships?

4.1   Data Collection and Dataset Characteristics

To answer the above research questions, we applied the relation discovery framework and monitored Twitter as well as news media over a period of two months starting from 15 November 2010. Overall, we crawled more than 10 million tweets and more than 70,000 news articles, which were published by prominent news Web sites such as BBC, CNN and New York Times. Figure 2(a) shows the number of Twitter posts and news articles published during the observation period. For news, the number of news articles published per day ranges between 100 and 1000 articles and follows rather characteristic patterns: on weekends the number of articles decreases clearly. Regarding Twitter posts, we see that the amount of tweets published each day is much higher and varies much stronger (between 50,000 and 400,000). However, we could not detect correlations between weekdays/weekends and the number of Twitter messages posted on those days. Two of the minima were caused by temporary unavailability of the Twitter monitoring service.

---

[5] http://dev.twitter.com/pages/streaming_api

[6] http://code.google.com/p/boilerpipe/

[7] http://www.opencalais.com/

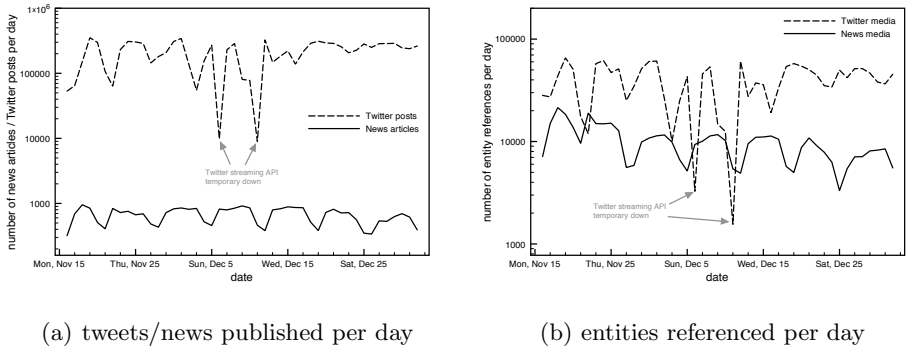(a) tweets/news published per day   (b) entities referenced per day

Fig. 2. Dataset characteristics: (a) the number of Twitter posts and news articles published per day and (b) the number of entity references identified per day

We processed all Twitter messages and news articles via our framework to extract entities that are mentioned in the tweets and news respectively. The number of entity references per day that we obtained by this extraction process is depicted in Figure 2(b). It is interesting to see that the amount of entity references observed in Twitter (approximately 10,000-100,000 references per day) is higher than for the News (approximately 5,000-20,000 references per day), but does not differ as strongly as for the number tweets/articles published each day. In fact, it is much harder to detect entities in short Twitter messages than news articles. For more than 40% of the processed tweets, we could not detect any entity while for 99.3% of the news articles we succeeded in detecting at least one entity. Hence, news articles seem to be a more valuable source for detecting entities than tweets. This observation gives our approach of exploiting news for detecting relations among entities mentioned in Twitter the first justification.

However, do the entities mentioned in Twitter overlap with those mentioned in news articles? To answer this question, we measured the overlap of entities that occur in both the Twitter media and the news media. For the given period, we observe that 72.6% of the top 1000 mentioned entities in Twitter are mentioned in the news media as well. This observation further validates the feasibility of considering news articles to detect relations between entities mentioned in Twitter and vice versa.

Figure 3 lists the number of distinct entities we observed in Twitter for the 39 different entity types. Persons, locations and organizations were mentioned most often, followed by movies, music albums, sport events and political events. In our analysis, we investigate the discovery of relationships between these entities. Moreover, we analyze specific types of relations such as relationships between persons and locations or organizations and events in detail. The complete list of those relation types that we analyzed in detail can be obtained via the supporting website [1].
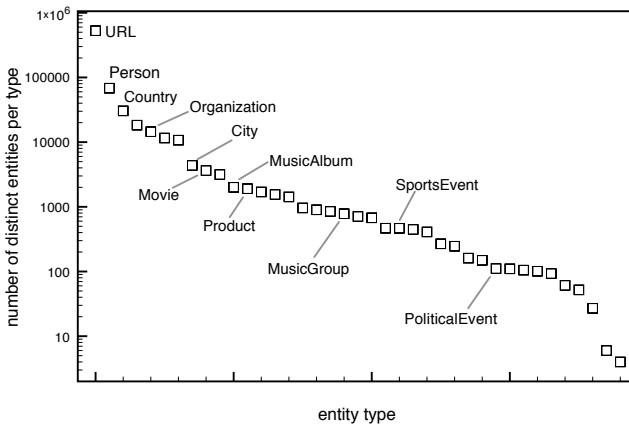
Fig. 3. Number of distinct entities per entity type mentioned in Twitter (39 different types)

4.2   Methodology, Metrics and Ground Truth

The relation detection task can be interpreted as a ranking problem. Given the dataset described above, the relation detection strategies have to rank entity pairs (of given types) so that those pairs, which are *truly* related (at a specific point in time) appear at the top of the ranking. We obtained the corresponding ground truth of *true* relationships (i) via DBpedia [5] and (ii) via a user study.

To utilize DBpedia—the structured representation of Wikipedia—as ground truth, we mapped entities to their corresponding DBpedia resources by exploiting the DBpedia lookup service[8]. We considered a given entity pair (e_1, e_2) as related if e_1 referenced or mentioned e_2 in one of its properties (e.g. in the full text description) or vice versa. However, for more than 35% of the entities we could not retrieve appropriate DBpedia URIs, for example, because these entites were not mentioned in DBpedia (yet). Therefore, we also conducted a user study to complement the DBpedia-based ground truth.

In our user study, we asked PhD students from our group to judge whether two entities, for which a relationship was detected by one of the tested strategies, are really related on a four-point scale: "not related or unknown", "rather not related", "rather related" and "related". Overall, we obtained 1294 judgments. 62.6% of the relationships were rated as "related" and were therefore considered as ground truth. We further asked the participants to judge whether a relationship is valid at a specific period in time. For example, whether they think that "Jullian Assange (person) and London (city)" or "Bill Clinton (person) and presidential election (political event)" are related during a certain time-frame (one week). Given such temporal constraints, 57.3% of the relations were classified as related.

Given the ground truth, we measured the quality of the relation discovery strategies by means of the following information retrieval metrics.
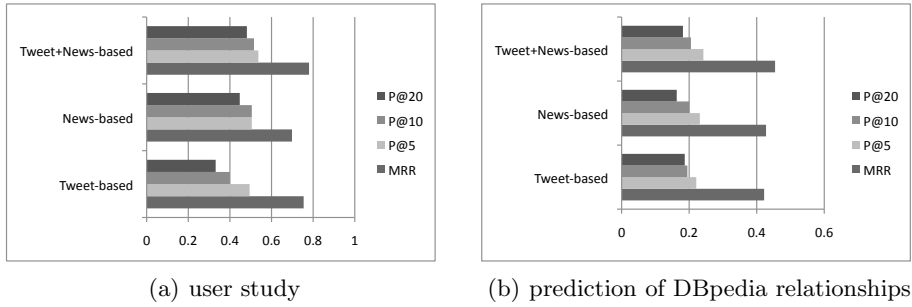
---

[8] http://dbpedia.org/lookup

(a) user study   (b) prediction of DBpedia relationships

Fig. 4. Accuracy of different relation discovery strategies based on (a) ground truth obtained from the user study and (b) ground truth obtained from DBpedia (more than 5,000 relations).

MRR. The *MRR* (Mean Reciprocal Rank) indicates at which rank the first *true* relationship occurs on average.

P@k. Precision at rank k (*P@k*) represents the average proportion of *true* relationships within the top k.

R@k. Recall at rank k (*R@k*) is the number of *true* relationships, which are listed within the top k, divided by the number of known *true* relationships.

**F@k.** F-measure at rank k (*F@k*) is the harmonic mean of R@k and P@k.

4.3   Results

To answer the first research question, we summarize the accuracies of the different strategies in Figure 4. For both test settings – user study and prediction of relationships in DBpedia – we see that by combining the Tweet-based and News-based strategies (Tweet+News-based), we achieve higher precision. In particular, the Tweet+News-based strategy performs better with respect to the mean reciprocal rank (MRR), P@5 and P@10. The results from the user study (Figure 4(a)) reveal that the Tweet-based strategy performs slightly better than the news-based strategy regarding MRR. However, the precision of the Tweet-based strategy decreases stronger than the precision of the News-based strategy when k is increased. In fact, this observation is influenced by the number of candidate relationships: all strategies are based on the co-occurrence of entities and, as tweets are limited to 140 characters in length, we observe that the probability of co-occurring entities is smaller in Twitter. For example, for relationships between *persons* and *political events* the probability that such entities co-occur in a tweet of our Twitter dataset is 0.00025, which results in 146 co-occurring persons-event candidate pairs. In contrast, the news corpus features a probability of 0.32 and 7893 person-event co-occurrences in total. Hence, the consideration of news promises to detect further relationships and therewith can potentially increase recall.

Figure 5 confirms this hypothesis. With respect to recall (R@10) and F-measure, the News-based strategy performs (32%) significantly better than the Tweet-based strategy (see Figure 5(a)). By exploiting both tweets and news, we
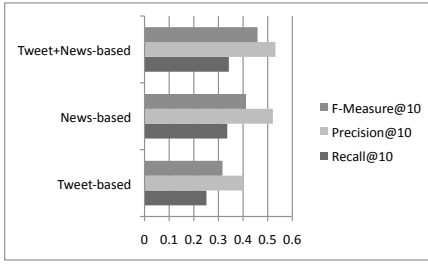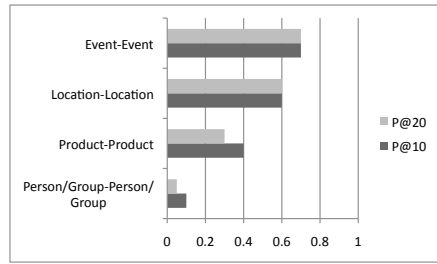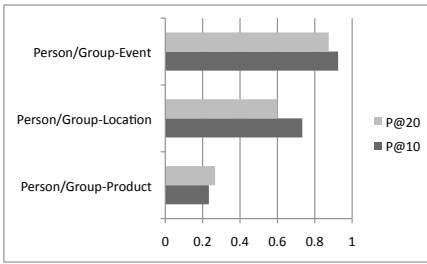
(a) Precision, Recall and F-measure
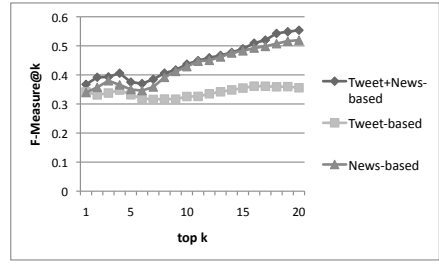
(b) F-measure within top k

Fig. 5. Precision and recall: (a) precision, recall and F-measure within the top 10 and (b) F-measure for different ranks


(a) Person/Group relations to events, locations and products.

(b) Relations between entities that have the same type.

Fig. 6. Accuracy of different types of relations: (a) relations of persons or groups with events, locations and products and (b) relations between entities that are of the same type

achieve even higher performances regarding recall (with 36%) and F-measure (with 45%). Figure 5(b) further illustrates how the F-measure performs at different ranking levels. While the F-measure of the Tweet-based strategy saturates quickly, the F-measure of the Tweet+News-based strategy increases and continuously achieves higher F-measure than the other strategies. We conclude that by combining both tweets and news articles, we achieve the best performance (regarding recall, precision and F-measure) in discovering typed relationships among entities – which answers the first research question raised at the beginning of this section.

Figure 6 depicts the accuracy of the Tweet+News-based strategy for specific types of relations and allows us to answer the second research question. To better visualize the results, we grouped the different types of relations into more abstract classes. For example, *Person/Group-Event* relationships cover relations between persons and political events, persons and sport events, organizations and sport events, et cetera. Figure 6(a) shows the precisions for inferring relationships between persons or groups and (i) events, (ii) locations (including cities or countries) and (iii) products (including movies or music albums). It is
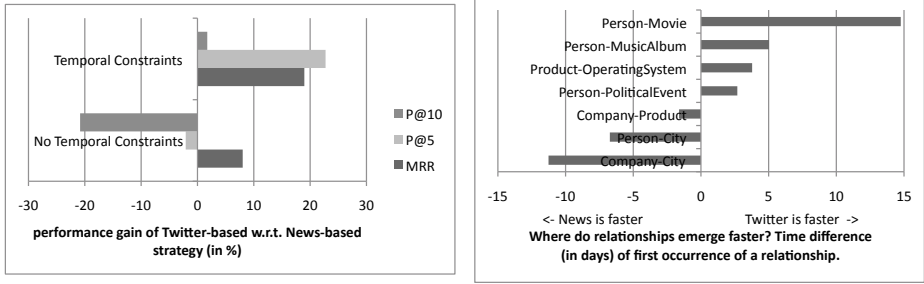
(a) Relations with(out) temporal constraints.

(b) Time difference in detecting relations.

Fig. 7. Temporal aspects: (a) difference in accuracy between news-based and tweet-based strategies for relations with/without temporal constraints and (b) time difference between news-based and tweet-based strategies in detecting certain relations

interesting to see that the Tweet+News-based strategy discovers relationships between persons/groups and events with higher precision—0.92 and 0.87 regarding P@10 and P@20—than people's relations to products (0.23 and 0.26) or locations (0.73 and 0.6). The good performance in the context of events can be explained by the nature of tweets and news articles, which are exploited to detect the relationships. Both media are centered around news events [16] which makes them a good source for inferring relationships between events and other entities.

Figure 6(b) shows that relationships between two events can also be discovered with high precision, followed by relations between locations (e.g. countries and cities, natural features and countries). For relationships among products (e.g. mobile devices and operating systems), the Tweet+News-based detection strategy performs better than for relationships between people, for which we observe the lowest performance (0.1 for P@10). However, in the context of relations between people, Twitter seems to be a much more reliable source. The Tweet-based strategy achieves, for example, a precision of 0.4 (P@10) for inferring relationships between persons and music groups while the News-based strategy allows for only 0.1 (P@10). One reason for this performance difference is the number of distinct entities that co-occur in a news article or Twitter message. On average, each news article contains 19.3 entities while a Twitter message allows for the identification of, on average, less than 1.8 entities. Intuitively, one can thus interpret these observations as follows: if two entities co-occur in a tweet then there is a high chance that there exists also a relationship between these entities (e.g. in DBpedia).

Some relationships have temporal constraints. Relations between persons and organizations or persons and locations might be valid or trending only for a certain period in time. For example, in December 2010, *London* and *Julian Assange*, the founder of WikiLeaks, were highly related, because Assange was under arrest in London at that time. The identification of such trending relationships could improve topic-based Twitter and news exploration interfaces. Figure 7(a) illustrates that Twitter is more appropriate for inferring relationships, which have

temporal constraints, than the news media. The Tweet-based strategy improves precision (P@5) by 22.7% in comparison with the News-based strategy, which in turn performs slightly better—with respect to P@5 and P@10—for detecting relationships without temporal constraints.

For the identification of trending relationships, it is further interesting to know which source of information allows for fast discovery of relationships. Figure 7(b) shows that particularly relationships between persons and movies or music albums emerge much faster (14.7 and 5.1 days respectively) in Twitter than in the traditional news media.

5   Discussion

Our results prove the feasibility of learning relationships between entities based on Twitter messages. However, we observed that the Tweet-based strategy achieves poorer results, with respect to the precision within the top k, when we increase the k value, while more stable behavior was noticed by the News-based strategy under the same conditions. On the other hand, the high MRR value for the Tweet-based strategy suggests that if two entities do co-occur in such a limited-size space, then there is a higher probability of strong correlation between them. This statement, however, does not apply to news articles where there are more identified entities on average than for tweets. Therefore, a strategy that combines the news and tweets to increase the probability of identifying entities, and then looking at their co-occurrence frequency to detect more relations, namely the Twitter+News-based strategy, takes the best out of the two in order to produce the highest performance (recall/precision) for discovering relations.

We described already one application of this relationship discovery in our motivations: content exploration in Twitter. When a user is reading a tweet, we can provide a list of entities that are related to the entities in the current tweet with the aid of our learnt relations. This allows users to follow a link in a chosen category (entity type) that they want to explore more, facilitating faceted browsing. For example, given a single Twitter message that mentions a person, learnt relationships link events, organizations or products related to this person and can thus be used to explore related content.

Having identified the best strategy and the media source for detecting relations between entities, we further investigated whether the accuracy varied depending on the entity and relation types. We observed that our Twitter+News-based strategy performs well in detecting relations between events and other types of entities, but not so well for relationships among persons. We see one possible explanation in the disambiguities in some of the typed entities, especially with synonymous entity values. For example, we observed cases where the same person was identified in different ways in the Person entity, such as "Barack Obama", "Obama", "Mr. Obama" and "President Obama". The co-occurrence-based strategies treat each of these representations as individual entities without merging them.

On the other hand, we can see from the results that relations among events and among locations can be discovered with high success. Such relations can be used to suggest further sources of information to a user who is reading up on an event in order to point them in the direction of related events. For example, a user viewing a tweet about the "Grammy Awards" to be held in Los Angeles on the 13th February might be interested to read about the "Academy Awards", which will happen later that month in the same city. Providing trending relations as a separate "list of suggestions" to the user would also enhance browsing means and the ability to access (interesting/desired) information when the information is still relevant. For this reason, we examined the quality and speed for discovering relationships with temporal dynamics in order to evaluate our strategies with media sources to eventually find trending relations. Tweet-based strategy was observed to perform highly in detecting relations with temporal validity both in terms of precision and speed when compared with the News-based strategy.

Our analysis of how fast we can discover relations further showed that the source of information first comes out in tweets, before it appears in the news especially for certain types of relations, e.g. Person-Movie, Person-MusicAlbum, Product-OperatingSystem and Person-PoliticalEvent. A possible explanation for Twitter's considerable speed and accuracy in finding "trending topics" over news media could be explained by comparing the number and frequency of tweets to that of the news articles posted, as was shown in Figure 2(a). Another explanation for this could be the fact that Twitter is used as a tool for fast communication, where news media is mostly for information broadcasting/reporting. Looking at the graph in Figure 2(a), one can see that there is more of a regular pattern for the news articles, while the number of tweets are more irregular, suggesting that they vary according to the current trend(s) more so than the day of the week. Therefore, there is no surprise that Tweet-based strategy has better performance in detecting relations with temporal validity. This also describes why relations types such as Company-Product, Person-City and Company-City are detected faster with news media than Twitter. These relations are long-term relations in general with less temporal constraints in the relation.

6   Conclusions

In this paper, we proposed a framework for discovering relationships between entities mentioned in microblog posts published on Twitter. Our framework extracts entities from Twitter messages, allows for further enrichment by exploiting traditional news media and provides different strategies for detecting typed relationships that moreover may have temporal constraints. Our analysis revealed that the consideration of news articles improves recall and precision of the relation discovery clearly. Furthermore, the accuracy depends strongly on the type of relationships that should be learnt: while relations among persons are fairly difficult to detect, relationships between persons and events can be discovered with high precision of more than 80%. For detecting relationships that have a limited temporal validity, Twitter-based strategies outperform News-based strategies by

more than 20% regarding precision (P@5). Further, by exploiting Twitter it is possible to discover certain types of trending relationships faster than by considering traditional news media.

Relations learnt by our relation discovery framework allow for supporting topic-based content exploration in Twitter and can be applied to suggest keywords when querying for Twitter content. Moreover, relationships can be exploited to enrich and complement existing ontologies, for example, by exploiting the linkage to DBpedia entities as provided by our framework. The relation discovery framework can also be used for engineering content recommendation applications. Studying the performance of these applications is part of our future work.

References

1. Abel, F., Celik, I.: Supporting website: datasets, further details and additional findings (2011), `http://wis.ewi.tudelft.nl/icwe2011/relation-learning/`
2. Abel, F., Gao, Q., Houben, G.J., Tao, K.: Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web. In: Antoniou, et al. (eds.) Extended Semantic Web Conference (ESWC), Heraklion, Greece, Springer, Heidelberg (2011)
3. Akcora, C.G., Bayir, M.A., Demirbas, M., Ferhatosmanoglu, H.: Identifying Breakpoints in Public Opinion. In: Melvile, P., Leskovec, J., Provost, F. (eds.) Proceedings of Workshop on Social Media Analytics (SOMA) at KDD 2010, Washington, DC, USA (2010)
4. Bernstein, M., Kairam, S., Suh, B., Hong, L., Chi, E.H.: A torrent of tweets: managing information overload in online social streams. In: Proceedings of the CHI Workshop on Microblogging: What and How Can We Learn From It? (2010)
5. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - A crystallization point for the Web of Data. In: Web Semantics: Science, Services and Agents on the World Wide Web (2009)
6. 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.) Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM). The AAAI Press, Washington, DC, USA (2010)
7. Chen, J., Nairn, R., Nelson, L., Bernstein, M., Chi, E.: Short and tweet: experiments on recommending content from information streams. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI), pp. 1185–1194. ACM, New York (2010)
8. 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: Proceedings of the 19th International Conference on World Wide Web (WWW), pp. 331–340. ACM, New York (2010)

---

[9] `http://imreal-project.eu`

9. Honeycutt, C., Herring, S.C.: Beyond microblogging: Conversation and collaboration via twitter. In: Proceedings of the 42nd Hawaii International Conference on Systems Science (HICSS), pp. 1–10. IEEE, Big Island (2009)

10. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Emergent Semantics in Bib-Sonomy. In: Hochberger, C., Liskowsky, R. (eds.) Informatik 2006: Informatik für Menschen. LNI, vol. 94(2). GI, Bonn (2006)

11. Huang, J., Thornton, K.M., Efthimiadis, E.N.: Conversational Tagging in Twitter. In: Chignell, M.H., Toms, E. (eds.) Proceedings of the 21st ACM Conference on Hypertext and Hypermedia (HT), pp. 173–178. ACM, New York (2010)

12. Hughes, A.L., Palen, L.: Twitter Adoption and Use in Mass Convergence and Emergency Events. In: Landgren, J., Jul, S. (eds.) Proceedings of the International Conference on Information Systems for Crisis Response and Management ISCRAM 2009 (May 2009)

13. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. WebKDD/SNA-KDD 2007, pp. 56–65. ACM, New York (2007)

14. Kaufman, S.J., Chen, J.: Where we Twitter. In: Proceedings of the CHI Workshop on Microblogging: What and How Can We Learn From It? (2010)

15. Kohlschütter, C., Fankhauser, P., Nejdl, W.: Boilerplate detection using shallow text features. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM), pp. 441–450. ACM, New York (2010)

16. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web (WWW 2010), pp. 591–600. ACM, New York (2010)

17. Lerman, K., Ghosh, R.: Information contagion: an empirical study of spread of news on digg and twitter social networks. In: Proceedings of 4th International Conference on Weblogs and Social Media, ICWSM (2010)

18. Marinho, L.B., Buza, K., Schmidt-Thieme, L.: Folksonomy-based collabulary learning. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T.W, Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 261–276. Springer, Heidelberg (2008)

19. Owens, J.W., Lenz, K., Speagle, S.: Trick or Tweet: How Usable is Twitter for First-Time Users? Usability News 11 (2009)

20. Romero, D.M., Meeder, B., Kleinberg, J.: Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In: Proceedings of the 20th International Conference on World Wide Web (WWW). ACM, New York (2011)

21. Rowe, M., Stankovic, M., Laublet, P.: Mapping Tweets to Conference Talks: A Goldmine for Semantics. In: Passant, A., Breslin, J., Fernandez, S., Bojars, U. (eds.) Workshop on Social Data on the Web (SDoW), Colocated with ISWC 2010, CEUR-WS.org, Shanghai, China, vol. 664 (2010)

22. Yardi, S., boyd, d.: Dynamic Debates: An Analysis of Group Polarization over Time on Twitter. Bulletin of Science, Technology and Society 30 (2010)

23. Zhao, D., Rosson, M.B.: How and why people Twitter: the role that micro-blogging plays in informal communication at work. In: Proceedings of the ACM International Conference on Supporting Group Work (GROUP), pp. 243–252. ACM, New York (2009)