Adaptive Faceted Search on Twitter

Ilknur Celik[1], Fabian Abel[1], Patrick Siehndel[2]

[1] Web Information Systems, Delft University of Technology
{celik,abel}@tudelft.nl
[2] L3S Research Center, Leibniz University Hannover, Germany
siehndel@l3s.de

Abstract. In the last few years, Twitter has become a powerful tool for publishing and discussing information. Yet, content exploration in Twitter requires substantial efforts and users often have to scan information streams by hand. In this paper, we approach this problem by means of faceted search. We propose strategies for inferring facets and facet values on Twitter by enriching the semantics of individual Twitter messages and present different methods, including personalized and context-adaptive methods, for making faceted search on Twitter more effective.

Key words: faceted search, twitter, semantic enrichment, adaptation

1 Introduction

Twitter is a Social Web phenomenon that is attracting interest from people all around the world for a variety of different purposes [1], such as consuming and propagating news [2], crisis management [3] or communication with other people [4]. Over the last few years, Twitter has shown an exponential growth and became the most popular microblogging site with several hundreds of millions of users and more than 50 million Twitter messages (tweets) per day[3]. Highly active users regularly receive thousands of tweets every day [5]. This information overload may cause users to get lost in the information network, become demotivated and frustrated. Finding your way around Twitter is indeed not very straightforward due to the lack of a user-friendly browsing option that goes beyond the existing chronologically-ordered clutter option [5, 6].

Recently, researchers started to study strategies for recommending URLs [7], news articles [8] or entire conversations on Twitter [9]. However, search on Twitter has not been studied extensively yet which motivates, for example, the TREC 2011 track on Microblogs that defines first search tasks on Twitter[4]. In line with the TREC research objectives, we investigate ways to enhance search and content exploration in the microblogosphere by means of faceted search.

Traditional faceted search interfaces allow users to search for items by specifying queries regarding different dimensions and properties of the items (facets) [10]. For example, online stores such as eBay[5] or Amazon[6] enable end-users to narrow

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

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

[5] http://ebay.com/

[6] http://amazon.com/

down their search for products by specifying constraints regarding facets such as the price, the category or the producer of a product. In contrast, information on Twitter is rather unstructured. Tweets are short text messages that do not explicitly feature facets. How can facets be extracted from tweets and how can we design appropriate faceted search strategies on Twitter? In this paper, we answer these questions and introduce an adaptive faceted search framework for Twitter. Our main contributions can be summarized as follows.

Semantic Enrichment We present methods for enriching the semantics of tweets by extracting facets (entities and topics) from tweets and related external Web resources.

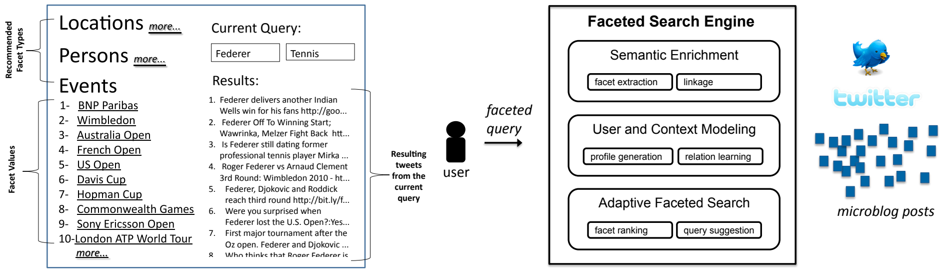
User and Context Modeling Given the semantically enriched tweets, we propose user and context modeling strategies that identify (current) interests of a given Twitter user and allow for contextualizing the demands of this user.

Adaptive Faceted Search We introduce faceted search strategies for content exploration on Twitter and propose methods that adapt to the interests and context of a user.

2 Faceted Search on Twitter

On Twitter, facets describe properties of a Twitter message. For example, persons that are mentioned in a tweet or events a tweet refers to. Oren et al. [10] formulate the problem of faceted search in RDF terminology. Given an RDF statement $(subject, predicate, object)$, the faceted search engine interprets (i) the subject as the actual resource that should be returned by the engine, (ii) the predicate as the facet type and (iii) the object as the facet value (restriction value). A faceted query (facet-value pair) that is sent to a faceted search engine thus consists of a predicate and an object. We follow this problem formulation proposed by Oren et al. [10] and interpret tweets as the actual resources the faceted search engine should return. If a tweet (subject) mentions an entity then the type of the entity is considered as facet type (predicate) and the actual identifier of the entity is considered as facet value (object). For example, given a tweet t that refers to the tennis player "Federer", the corresponding URI of the entity $(URI_{federer})$ and the URI of the entity type (URI_{person}) are used to describe the tweet by means of an RDF statement: $(t, URI_{person}, URI_{federer})$.

Figure 1(a) illustrates how we envision the corresponding faceted search interface that allows users to formulate faceted queries. Given a list of facet values which are grouped around facet types such as locations, persons and events, users can select facet-value pairs such as $(URI_{event}, URI_{wimbeldon})$ to refine their current query (see top right in Fig. 1(a): $(URI_{person}, URI_{federer})$, $(URI_{sportsgame}, URI_{tennis})$). A faceted query thus may consist of several facet-value pairs. Only those tweets that match all facet-value constraints will be returned to the user. The ranking of the tweets that match a faceted query is a research problem of its own and could be solved by exploiting the popularity of tweets – e.g. measured via the number of re-tweets or via the popularity of the user who published the tweet (cf. [11]). The core challenge of the faceted search interface is to support the facet-value selection as good as possible. Hence, the


Recommended Facet Types
Locations more...
Persons more...
Events
Facet Values
1- BNP Paribas
2- Wimbledon
3- Australia Open
4- French Open
5- US Open
6- Davis Cup
7- Hopman Cup
8- Commonwealth Games
9- Sony Ericsson Open
10-London ATP World Tour
more...
Current Query:
Federer
Tennis
Results:
1. Federer delivers another Indian Wells win for his fans http://goo...
2. Federer Off To Winning Start; Wawrinka, Melzer Fight Back htt...
3. Is Federer still dating former professional tennis player Mirka ...
4. Roger Federer vs Arnaud Clement 3rd Round: Wimbledon 2010 - ht...
5. Federer, Djokovic and Roddick reach third round http://bit.ly/f...
6. Were you surprised when Federer lost the U.S. Open?:Yes...
7. First major tournament after the Oz open. Federer and Djokovic ...
8. Who thinks that Roger Federer is
Resulting tweets from the current query
user
faceted query
Faceted Search Engine
Semantic Enrichment
facet extraction
linkage
User and Context Modeling
profile generation
relation learning
Adaptive Faceted Search
facet ranking
query suggestion
twitter
microblog posts

(a) Faceted search interface (b) Faceted search architecture

Fig. 1. Adaptive faceted search on Twitter: (a) example interface and (b) architecture of the faceted search engine.

facet-value pairs that are presented in the faceted search interface (see left in Figure 1(a)) have to be ranked so that users can quickly narrow down the search result lists until they find the tweets they are interested in. Therefore, the *facet ranking problem* can be defined as follows.

Definition 1 (Facet Ranking Problem). *Given the current query F_{query}, which is a set of facet-value pairs $(predicate, object) \in F_{query}$, the hit list H of resources that match the current query, a set of candidate facet-value pairs $(predicate, object) \in F$ and a user u, who is searching for a resource t via the faceted search interface, the core challenge of the faceted search engine is to rank the facet-value pairs F. Those pairs should appear at the top of the ranking that restrict the hit list H so that u can retrieve t with the least possible effort.*

The effort, which u has to invest to narrow down the search result list H, can be measured by click and scroll operations (e.g. the number of facet-value pair selections). Our goal is to provide a faceted search interface that minimizes the effort a user has to invest in retrieving the tweets in which a user is interested in. Query suggestions and ranking facet-value pairs according to the current demands of a user are therefore essential and will be discussed in the next section.

2.1 Architecture for Adaptive Faceted Search on Twitter

Figure 1(b) illustrates the architecture of the engine that we propose for faceted search on Twitter. The main components of the engine are the following.

Semantic Enrichment The semantic enrichment layer aims to extract facets from tweets and generate RDF statements that describe the facet-value pairs which are associated with a Twitter message. In particular, each tweet is processed to identify entities (facet values) that are mentioned in the message. We therefore make use of the OpenCalais API[7], which allows for the extraction of 39 different types of entities (facet types) including persons, organizations, countries, cities and events. As Twitter messages are limited to 140 characters, the extraction of entities from tweets is a non-trivial problem. Thus, we introduced a set of strategies that link tweets with external Web resources (news articles) and propagate the semantics extracted from these resources to the related tweets

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

in [8]. For example, given a tweet “This is great http://bit.ly/2fRds1t”, we extract entities from the referenced resource (http://bit.ly/2fRds1t) and attach the extracted entities to the tweet.

User and Context Modeling In order to adapt the facet ranking to the people who are using the faceted search engine, we propose user modeling and context modeling strategies. The user modeling strategies model the interests of the users in certain facet values (entities and topics). We therefore exploit the tweets that have been published (including re-tweets) by a user. In future work, we also plan to consider click-through data from the faceted search engine. Context modeling covers mining of new knowledge from the Twitter data. We therefore propose relation learning strategies that exploit co-occurrence of entities in Twitter messages to infer typed relationships between entities [12].

Adaptive Faceted Search Based on the semantically enriched tweets, the learnt relationships between entities extracted from tweets and the user profiles generated by the user modeling layer, the adaptive faceted search layer solves the actual facet ranking problem. It provides methods that adapt the facet-value pair ranking to the given context and user. Furthermore, it provides query suggestions by exploiting the relations learnt from the Twitter messages. Given the current facet query, which is a list of facet-value pairs where each value refers to an entity, we can exploit relationships between entities in order to identify entities that are related to those entities that occur in the current facet query. We leave the analysis of such query suggestions for future work. Instead, we focus on the facet ranking problem and propose different strategies for ranking facet-value pairs in the next subsection.

2.2 Adaptive Faceted Search and Facet Ranking Strategies

Non-Personalized Facet Ranking A lightweight approach is to rank the facet-value pairs $(p, e) \in F$ based on their occurrence frequency in the current hit list H, the set of tweets that match the current query (cf. Definition 1):

$$rank_{frequency}((p, e), H) = |H_{(p,e)}| \tag{1}$$

$|H_{(p,e)}|$ is the number of (remaining) tweets that contain the facet-value pair (p, e) that can be applied to further filter the given hit list H. By ranking those facets that appear in most of the tweets, $rank_{frequency}$ minimizes the risk of filtering out relevant tweets but might increase the effort a user has to invest to narrow down search results.

Context-adaptive Facet Ranking The context-adaptive strategy exploits relationships between entities (facet values) to produce the facet ranking. A relationship is therefore defined as follows:

Definition 2 (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. We use co-occurrence frequency as weighting scheme. Hence, given the enriched tweets, we count the number of tweets both entities (e_1 and e_2) are associated with. The context-adaptive facet ranking strategy ranks the facet-value pairs $(p, e) \in F$ according to $w(e_i, e)$, where e_i is a facet value that is already part of the given query: $(p_i, e_i) \in F_{query}$ (cf. Definition 1):

$$rank_{relation}((p, e), F_{query}) = \sum_i w(e_i, e)|(p, e_i) \in F_{query} \tag{2}$$

Hence, the context-sensitive strategy can only be applied in situations where the user has already made one selection, so that $|F_{query}| > 0$.

Personalized Facet Ranking The personalized facet ranking strategy adapts the facet ranking to a given user profile that is generated by the user modeling layer depicted in Figure 1(b). User profiles conform to the following model and specify a user's interest into a specific facet value (entity).

Definition 3 (User Profile). *The profile of a user $u \in U$ is a set of weighted entities where with respect to the given user u for an entity $e \in E$ its weight $w(u, e)$ is computed by a certain function w.*

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

Here, E and U denote the set of entities and users respectively.

Given the set of facet-value pairs $(p, e) \in F$ (see Definition 1), the personalized facet ranking strategy utilizes the weight $w(u, e)$ in $P(u)$ to rank the facet-value pairs:

$$rank_{personalized}((p, e), P(u)) = \begin{cases} w(u, e) & \text{if } w(u, e) \in P(u) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

By combining the above three strategies it is possible to generate further facet ranking methods. A combination of two strategies can be realized by building the weighted average computed for a given facet-value pair (p, e) (e.g. $rank_{combined} = \alpha \cdot rank_{\alpha}((p, e)) + (1 - \alpha) \cdot rank_{\beta}((p, e))$, where $\alpha \in [0..1]$).

3 Preliminary Analysis and Future Work

In this paper, we explored strategies for faceted search on Twitter. We presented a framework that enriches the semantics of Twitter messages in order to generate facets that describe the content of tweets (e.g. persons, locations, organizations). To do so, we extracted entities both from tweets and linked external Web resources. Our analysis based on a large Twitter dataset[8] showed that the exploitation of links for enriching tweets is necessary to better support faceted search on Twitter, due to the obvious increase in the number of facet values when tweets are enriched with entities of related news articles. We proposed an adaptive faceted search engine with different strategies for ranking facets including methods that adapt to the actual context and user. The context-adaptive method exploits relationships between facet values (entities) that we learn from

[8] We make our dataset publicly available at `http://wis.ewi.tudelft.nl/umap2011/`

tweets and linked news articles [12]. Our strategy discovers relationships between persons/groups (including organizations) and events (including political, sports and entertainment) with high precision of 0.92 and 0.87 regarding P@10 and P@20. Our analysis indicates that these relationships can be used to suggest and rank facets that are related to the current context (the current faceted query) with high precision.

Furthermore, the personalized facet ranking requires a user profile $P(u)$ in order to adapt the facet ranking to the preferences of the user. In [8], we showed that our user modeling strategies, which are based on semantic enrichment of tweets, outperform other strategies such as hashtag-based approaches significantly for recommending news articles. In future work, we will analyze the applicability of those user modeling strategies for our adaptive faceted search engine. In our evaluation, we will measure the effect of our facet ranking strategies (i) in the context of an automatic experiment as proposed by Koren et al. [13] and (ii) in practice by enabling real users to experiment with our adaptive faceted search interface for Twitter.

References

1. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: WebKDD/SNA-KDD, ACM (2007) 56–65
2. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: WWW, ACM (2010) 591–600
3. Hughes, A.L., Palen, L.: Twitter Adoption and Use in Mass Convergence and Emergency Events. In: ISCRAM, iscram.org (2009)
4. Zhao, D., Rosson, M.B.: How and why people Twitter: the role that micro-blogging plays in informal communication at work. In: GROUP, ACM (2009) 243–252
5. Bernstein, M., Kairam, S., Suh, B., Hong, L., Chi, E.H.: A torrent of tweets: managing information overload in online social streams. In: CHI Workshop on Microblogging (2010)
6. Owens, J.W., Lenz, K., Speagle, S.: Trick or Tweet: How Usable is Twitter for First-Time Users? Usability News **11** (2009)
7. Chen, J., Nairn, R., Nelson, L., Bernstein, M., Chi, E.: Short and tweet: experiments on recommending content from information streams. In: CHI, ACM (2010) 1185–1194
8. Abel, F., Gao, Q., Houben, G.J., Tao, K.: Analyzing User Modeling on Twitter for Personalized News Recommendations. In: UMAP, Springer (2011)
9. Chen, J., Nairn, R., Chi, E.H.: Speak Little and Well: Recommending Conversations in Online Social Streams. In: CHI, ACM (2011)
10. Oren, E., Delbru, R., Decker, S.: Extending faceted navigation for RDF data. In: ISWC, Springer (2006) 559–572
11. Weng, J., Lim, E.P., Jiang, J., He, Q.: TwitterRank: finding topic-sensitive influential Twitterers. In: WSDM, ACM (2010) 261–270
12. Celik, I., Abel, F., Houben, G.-J.: Learning Semantic Relationships between Entities in Twitter. In: ICWE, Springer (2011)
13. Koren, J., Zhang, Y., Liu, X.: Personalized interactive faceted search. In: WWW, ACM (2008) 477–486