Abstract. In this paper we report about our current and ongoing research efforts aiming at knowledge discovery, offline social data mining and social entity extraction based upon semantic technologies. Further we are aiming to provide the scientific architecture paradigm for building semantic applications that rely on social data. In this early stage our work focuses on data from Twitter\(^1\) as currently most popular and fastest growing microblogging platform. In the realm of our research we implemented applications like Grabeeter\(^2\) for storing and caching the social data and STAT infrastructure that uses semantic standards like RDF (SIOC, FOAF), SPARQL and existing semantic services as Sindice\(^3\) and Linked Data silos as DBPedia\(^4\) or GeoNames\(^5\) as well. They represent parts of novel architecture paradigm for semantic social applications intended to be introduced here.

\textbf{Keywords:} Twitter, microblogging, semantic, linked data, SIOC, FOAF, SMOB, web 2.0, data portability, knowledge discovery, information retrieval, data mining, sentiment classification, social web

1 Introduction

Most of the content on the Web nowadays is user generated. Especially microblogging gained strong importance in recent years. Current microblogging

\(^1\)http://www.twitter.com/ (last access: September 2010)
\(^2\)http://grabeeter.tugraz.at/ (last access: September 2010)
\(^3\)http://www.sindice.com/ (last access: September 2010)
\(^4\)http://www.dbpedia.org/ (last access: September 2010)
\(^5\)http://geonames.org (last access: September 2010)
platforms like Twitter, Jaiku\textsuperscript{6}, Tumblr\textsuperscript{7}, Emote.in\textsuperscript{8} attract every day masses of users with different social, cultural and educational background. Furthermore microblogs tend to become a solid media for simplified collaborative communication. Templeton [3] defines microblogging as a small-scale form of blogging made up from short, succinct messages, used by both consumers and businesses to share news, post status updates and carry on conversations.

While communicating people share different kind of information like common knowledge, opinions, emotions, and information resources and their likes or dislikes. They discuss on different topics in more or less open communities [4]. Information chunks (small scale messages) delivered by masses and structured in adequate way offers new scientific insights which are interesting not only for science but also for commercial use like monitoring of trends, advertising, statistics, reputation management and news broadcasting. Possibility of monitoring and analysis of such messages not only restricted to humans but also for machines would enormously contribute the exploitation of already present on-line social data essence e.g. for educational, commercial or informative purposes. Short form of content posted by microbloggers offers also a solid base for automated content processing and analysis. However current state of research lacks of simplified architectural paradigm approach according to these problems.

Twitter along with Facebook\textsuperscript{9} belongs to the fastest growing social web platforms of the last 12 months\textsuperscript{10} [2]. On 22nd of February 2010 Twitter hits 50 million tweets per day\textsuperscript{11}. Without any exaggeration it can be said that these two social networks are worth to be researched in more detail [1].

This publication deals first with the question, what can be the advantages of an (web)-application that can also be used offline or online for information retrieval and semantically based knowledge discovery in a micro-content system like Twitter. Further we want to answer in which way a social data can be exposed using Semantic Web technologies in order to provide a solid interface for data mining, sentiment extraction, and knowledge discovery in general and as fundament for building smart semantic applications based upon social data as well. Finally as answer to these questions we would like to introduce our novel architectural paradigm approach for this research issue.

\textsuperscript{6} http://www.jaiku.com/ (last access September 2010)
\textsuperscript{7} http://www.tumblr.com/ (last access September 2010)
\textsuperscript{8} http://www.emote.in/ (last access: September 2010)
\textsuperscript{9} http://www.facebook.com/ (last access: September 2010)
\textsuperscript{10} http://ibo.posterous.com/aktuelle-twitter-zahlen-als-info-grafik (last access: 2010-04)
\textsuperscript{11} http://mashable.com/2010/02/22/twitter-50-million-tweets/ (last access: 2010-04)
2 Related Work

2.1 Microblogging and Twitter

Although the beginning of first serious microblogs dated back five years ago their leverage on the web grows rapidly. Especially in the area of communication and social networking microblogging is gaining significantly daily [5]. Tumblr, Jaiku, Emote.in and identi.ca are only some of current massively used microblogging platforms. Most significant among them is Twitter, which induced a new culture of communication [6] [7]. The 140 characters restriction for twitter messages can also be compared with a short-message internet-based service platform. Users can send a post (tweet) that is listed on the top of their twitter-wall together with messages of people they are following. Furthermore any user can be followed by anyone who is interested in that user’s updates. By nature Twitter or similar services support the fast exchange of different resources (links, pictures, thoughts) as well as fast and easy communication amongst more or less open communities [2]. In the same way Java et al. [11] defined four main user behaviors why people are using Twitter - for daily chats, for conversation, for sharing information and for reporting news. It has been shown lately that usage of Twitter at conferences helps to increase reports, statements, and announcements as well as supports fast conversation between participants. Nowadays very often so called Twitter-streams done by hash tag search nearby the projection of an ongoing presentation [8] or placed at any other location at the conference support the conference administration, organization, discussions or knowledge exchange. From this point of view microblogging becomes a valuable service reported by different publications [8] [9].

The rules are simple and platform is easy to use. This approach through its simplicity and easy technical accessibility produces masses of information. Analysis of such content in a systematic manner using standards and adequate automated techniques enables the consumer of information to get valuable data categorized by aspects like locality, time, popularity, categorization etc. Although Twitter already has an API with advanced search functionality, retrieved data lacks of usability. Bringing these results into structured form with appropriate domain description using wide accepted vocabularies for a specific knowledge domain would increase the relevance of information retrieved through mining and exploration of such content. Second disadvantage of Twitter API is that results are restricted to the last 3200 tweets.

2.2 Semantic for Microblogging

User posts, relations and communication between them represent the essence of microblogging idea. The Communication paradigm presented in most microblog platforms like Twitter could be covered through following cases: User “A” says something on topic “T”, User “A” communicates with user “B”, User “A” broadcasts

12 http://apiwiki.twitter.com/ (last access: September 2010)
some content without addressing. All these communicational patterns can be easily mapped into a tripartite structure [20][21], which mainly corresponds to the basic idea of RDF Framework.

Regarding Twitter there have already been done some efforts like Semantic Tweet\(^\text{13}\) to bring data from Twitter user into a wise semantic form. In order to describe semantically circumstances of relations between the users widely used FOAF (Frien-Of-A-Friend) vocabulary [10] is recommended to use and it will be considered by our architecture paradigm. For the purposes of posts’ description and relations around them like topic, author, content and the vocabulary called SIOC (Semantically Interlinked Online Communities)[17] is provided by Semantic Web Community. Currently a well known scientific project dealing with the semantic realization of microblogging platforms named SMOB (Semantic MicrOBlogging) [11] [12] is maintained by DERI Galway. It provides a SPARQL\(^\text{14}\) API and relies on vocabularies like FOAF, SIOC and OPO (Online Presence Ontology). Additionally it offers the interfaces to the semantic search engines like Sindice\(^\text{15}\) and to the Linked Data Cloud (LOD)\(^\text{16}\). There is also Smesher\(^\text{17}\) a micro-blogging client maintained by Benjamin Nowack with the ability of local storage. It provides also a SPARQL API as part of its implementation.

According to the mentioned aspects in this section there are some proven domain vocabularies as FOAF, SIOC or semantic technologies like SPARQL provided by the Semantic Web Community, which can be used for triplification and semantic description of microblogs like Twitter.

### 2.3 Linked Data

The idea of Linked Data [14] and growing activities in this field cumulated into LOD Cloud (Linking Open Data Cloud) [18] offering in the meanwhile reliable semantic data sets like e.g. DBPedia [13], which is a semantic version of Wikipedia as reasonable mapping source. DBPedia team operates besides SPARQL endpoints also a lookup service\(^\text{18}\). For identifying people, organizations or locations there are some other verified linked data sets present in LOD Cloud like New York Times\(^\text{19}\), Geonames\(^\text{20}\) etc. Similar data sets also exist about music, books, movies, science etc. Linking semantic sources using simple principles described in [14] [15] turns the web into a large database not only available for human but also to intelligent agents [16].

Bringing Twitter data into this infrastructure would increase the relevance of its content and offer more profound information on social aspects of posted content [17].

---

\(^{13}\) [http://semantictweet.com/](http://semantictweet.com/) (last access: September 2010)

\(^{14}\) [http://www.w3.org/TR/rdf-sparql-query/](http://www.w3.org/TR/rdf-sparql-query/) (last access: September 2010)

\(^{15}\) [http://www.sindice.com](http://www.sindice.com) (last access: September 2010)

\(^{16}\) [http://richard.cyganiak.de/2007/10/1od/](http://richard.cyganiak.de/2007/10/1od/) (last access: September 2010)

\(^{17}\) [http://smesher.org/](http://smesher.org/) (last access: September 2010)

\(^{18}\) [http://lookup.dbpedia.org/](http://lookup.dbpedia.org/) (last access: September 2010)

\(^{19}\) [http://data.nytimes.com/](http://data.nytimes.com/) (last access: September 2010)

\(^{20}\) [http://www.geonames.org/](http://www.geonames.org/) (last access: September 2010)
Our current and future efforts are also inspired by this vision. In following chapter we are introducing architecture suitable to meet the intended requirements and expectations.

3 System Architecture

3.1 Overview

Basically our experimental system paradigm depicted in Fig. 1 describes the basic mining architecture consisting of three fundamental parts: data acquisition, data extraction, and analysis as final step.

In the acquisition phase tweets are simply grabbed and subsequently stored into a local database or passed automatically through to extraction phase. Extraction phase triplifies the microblogs content into RDF triples and stores the data into a triple store of choice. Also interlinking of tweet bodies is done in this step. Finally in the analysis phase stored RDF triples are exposed over SPARQL Endpoints or by using the Lookup Services to humans and machines.
3.2 Data Acquisition

3.2.1 Grabeeter

Before creating semantic triples data must be at first retrieved from the twitter server. Our main module that serves data acquisition is called Grabeeter. It simply grabs Twitter user timelines using the Twitter API. Every user who owns a Twitter account can initialize and grab his/her data with Grabeeter. The architecture of Grabeeter (see Fig. 2) consists of two main parts. The first part is a web application and second part of Grabeeter consists of a client application developed in JavaFX6 technology for accessing the stored information on a client side.

![Architecture of Grabeeter](image)

The Grabeeter web application (see Fig. 3) uses the Twitter API to retrieve tweets of predefined users. Tweets are stored in the Grabeeter database and on the file system as Apache Lucene21 index. In order to ensure an efficient search tweets must be indexed. The Grabeeter web application provides access to the Grabeeter database through its own REST style [19] API. This enables client applications to retrieve tweets and user information in an easy way by implementing this API. In contrast to the Twitter API Grabeeter API provides all stored tweets and makes no restriction over time.

21 [http://lucene.apache.org/java/docs/](http://lucene.apache.org/java/docs/) (last access: September 2010)
The Grabeeter client application (Fig 4.) is developed using JavaFX in order to be independent from different operating systems as well as to provide an easy upgrade process for the client application using Java Web Start. Furthermore it simplifies the way to store the retrieved tweets on the user’s local file system for later offline processing.

http://java.sun.com/javase/6/docs/technotes/guides/javaws/index.html (last access: September 2010)
3.3 Data Extraction

3.3.1 Triplification Module

Triplification module resides on the achievement of SMOB (Semantic MicrOBlogging) project. SMOB is a Semantic MicroBlogging framework based on state-of-the-art Semantic Web and Linked Data technologies like SIOC PHP API\(^23\) or ARC2\(^24\) interfaces. Some parts of framework are adapted into our TriplificationModule and used to triplify the user data and tweets, which were harvested by Grabeeter using the SIOC and FOAF vocabulary in similar manner as it was made in SMOB Project. Once the data is triplified it will be stored into a RDF Store. For experimental purposes we are using ARC2 store hence our implementation resides on PHP. Additionally the SMOB project also offers an interlinking feature using the MOAT\(^25\) and CommonTag\(^26\) vocabularies and interface connection to

\(^23\) [http://wiki.sioc-project.org/index.php/PHPExportAPI](http://wiki.sioc-project.org/index.php/PHPExportAPI) (last access: September 2010)

\(^24\) [http://arc.semsol.org/](http://arc.semsol.org/) (last access: September 2010)

\(^25\) [http://moat-project.org/](http://moat-project.org/) (last access: September 2010)

\(^26\) [http://commontag.org/](http://commontag.org/) (last access: September 2010)
Sindice, which will be also considered to be used to expose the harvested data. Currently the module implementation status is limited to simple triplification and storing of triplified data.

### 3.3.2 Interlinking Module

Once the content is semantically structured and described, content fields like post body, channel (topic) inside the post body, author or location can much easier explored and analyzed. SPARQL for example offers a possibility to use regular expressions on content, which were previously semantically described and structured. Analyzed content in this form is qualified for more accurate specification by mapping and interlinking the detected entities into verified data sources.

Our interlinking module is still in development. Currently it uses a simple interlinking algorithm based on regular expressions inside SPARQL queries over the fields of sioc:MicroblogPost (sioc:content, sioc:topic, foaf:name, foaf:maker). In following pseudo code a sample SPARQL query searches for all tweets including #vienna hastag is shown.

```
SELECT ?post ?content ?maker ?name
WHERE {
  ?post rdf:type sioct:MicroblogPost;
  foaf:maker ?maker;
  ?maker foaf:name ?name;
  sioc:content ?content.
  FILTER(regex(?content,#vienna))
}
```

Those queries are used for RDF store of tweets, which are grabbed by Grabeeter, triplified and stored in ARC2 triple store. At this stage of implementation we are trying to extract some more concrete entities like persons, locations and interlink them with DBPedia or GeoNames. In this way additional reliable data is semantically linked to social data artifacts enhancing them with proven references.

### 3.4 Analysis

Analysis is done using Semantic Technologies and simple text based techniques. Other very significant contribution of analysis phase includes the exposure of triplified raw microblog data using SPARQL endpoints or lookup services. This aspect offers a solid base for further exploitation of grabbed data by intelligent
semantic web applications. Especially for Twitter we developed for demonstration purposes so called STAT (Semantic Tweeter Analysis Tool) Infrastructure. STAT is still in development at this moment. However we are aiming creating an analysis system that would be able to answer simple questions about persons and their actions and interactions on microblogging platforms. A snapshot of current stage of STAT is depicted in Fig. 5.

4 Conclusion and Future Work

Grabeeter was launched in May 2010. The web application as well as the JavaFX client can be accessed at http://grabeeter.tugraz.at. Grabeeter grabs the timeline data from Twitter and on the other hand it provides longer lasting storage for tweets that would be lost due the 3200 tweets restriction of Twitter API. The rapid improvements in the mobile technology have led to an ascending trend of using mobile applications in recent years. Consequently more users use mobile devices to access online applications. It is planned to build the Grabeeter client as a mobile application for different platforms (Android, iPhone, JavaFX devices …). The adaptations, which must be performed, are mainly the view adjustment and an appropriate look and feel for the mobile environments.

Furthermore we consider involving more complex text processing techniques as well as more advanced approaches like simple NLP (Natural Language Processing) methods in order to make the entity extraction inside the Interlinking Module more accurate and widely applicable.
In this paper we introduced the interesting aspects about microblogs, how far they correspond with ideas from other research areas like Semantic Web or Linked Data. We also tried to answer how far those two areas can be combined to gain more knowledge and mine usable data out of social context of microblogs. We also outlined the importance and relevance of such or similar efforts by examples and arguments from current research. Through implementation of Grabeeter and STAT along with integration of open source projects like SMOB or technologies like SPARQL into our Triplification and Interlinking Module we approved that it is possible to grab, triplify and expose the social data from microblogs’ to intelligent semantic clients. Although some research questions have to be solved we are convinced that in this way prepared social data especially from microblogs will be interesting as base for future semantic web applications. We have also carried out that current state-of-the-art-technologies can support this idea. Finally we presented our architectural paradigm approach that delivers the answer to specified research issue.

5 References


