Documenting Contemporary Society by Preserving Relevant Information from Twitter

16
CHAPTER

Thomas Risse, Wim Peters,
Pierre Senellart, and Diana Maynard

can we preserve tweets and their contexts and still understand what they mean in 10+ years? #archive

WHY ARCHIVE TWITTER?

In recent years, Twitter has changed from a medium for posting personal updates or status information to a channel for sharing and distributing information of all kinds. Its increasingly pervasive nature is encouraging more and more people to give insights into their daily life and to stay in contact with friends. This also attracts many companies and media agencies, attempting to establish a more or less constant flow of information to their customers. The limitation to 140 characters reduces efforts, and focusses the tweet on the core information. The ease of use of Twitter and its availability on every smartphone also encourages people to act as citizen journalists and immediately report the events they witness. Twitter can thus be seen as the foremost channel for "breaking news", where information about events appears before being distributed via traditional channels. Follow-up messages on Twitter complement news articles with sentiments, opinions, and related information. Prominent examples for citizen journalism are the Arab Spring (Mourtada & Salem, 2011) or the "Miracle on the Hudson" ("Twitter First Off the Mark with Hudson Plane Crash Coverage", 2009).

As a side effect of its active and pervasive usage, Twitter documents contemporary society in rich detail. Tweets give valuable insights into individuals, groups, and organisations, and enable an understanding of the public perception of events, people, products, or companies, including the flow of information. While in the past reports about society were written by individuals and were therefore biased, today Twitter and other Social Web applications create the possibility of a live documentation of our society. It gives unprecedentedly rich and detailed insights into the day-to-day process of public communication. This will allow later generations to understand how topics were spreading, how sentiments and opinions were developing, or to better understand the impact of technological developments like Twitter on the evolution of culture and society as it is possible today.

The long-term preservation of public Twitter content and its accessibility for research is thus becoming a cultural necessity. For short-term usage, the probability that Twitter content remains accessible at Twitter itself can be assumed to be high. In the long-term perspective—meaning more than 10 years—no prediction can be made, as the experience with past popular Internet sites such as GeoCities shows. GeoCities—founded in 1994, bought by Yahoo! in 1999, and closed down in 2009—was a popular Web service for hosting free user homepages. Nowadays, some of these homepages are preserved in a Web archive, thanks to some last-minute crawling activities, while others are lost forever. To avoid such a loss of valuable information for Twitter, capturing its content and preserving it for future generations is necessary.

The aim of the capturing effort for Twitter should be to preserve the content, the presentation, and the social context scope of a tweet. According to Middleton (2012), "social context scoping is a critically important scope because it collects the subject alongside the social commentary for a more complete historical record". The U.S. Library of Congress (LoC) is currently archiving all tweets since Twitter's inception in 2006, but their accessibility is unclear (see also Chapter 13 by Zimmer & Proferes, in this volume). While on the one hand this archive is already a big achievement, on the other hand the access limitations constrain its usability. In addition, the LoC archive only holds the tweets, but not necessarily their social context.

Contextualised Twitter capturing goes beyond the pure collection of tweets. The limitation of 140 characters per tweet forces the poster to be very focussed and brief. Hence, there is little or no room for any introduction or explanation to help understand the tweet. To assist the reader of a tweet in the future, it is necessary to give them more information about its context. The context within Twitter is defined by the person who tweets, the topic defined by the hashtag (if one is present), but also by the answering and re-tweeting chain in which a tweet may participate. In addition, some tweets have links to external pages, which can provide more details about the topic of a tweet. An interesting application that highlights this requirement is Speak2Tweet (Speak2Tweet, 2012), set up in January 2011 during the Egyptian revolution, which allows the tweeting of a URL to voice recordings for those without an Internet connection by making a phone call to a designated number. On the one hand, to simply capture the tweets which contain such links might lead to a loss of highly valuable information about the Arab Spring. On the other hand, following links present in a tweet, and gathering other tweets with the same hashtag or from users @mentioned within a tweet allows preserving a more comprehensive context of the tweet.

For implementing the described contextualised crawl strategy, the European-funded project ARCOMEM (ARchive Communities' MEMories) (ARCOMEM, 2012) follows a two-step philosophy. In the first step, Twitter content is captured and analysed to extract semantic and contextual information. This information triggers in a second step the Web crawler to collect relevant content from the Web.

During the capturing of tweets and the crawling of their context, future usage should be taken into account. Bearing in mind the large number of tweets and related pages generated per day, efficient access mechanisms are mandatory. One means of going beyond standard, full-text search is to enrich each tweet with descriptive meta-information. This meta-information consists of (a) information directly gathered from Twitter (e.g., user, creation date, geolocation) and (b) information extracted from the content, such as topics, events, and sentiments. This meta-information can be used in conjunction with a full-text search to select the appropriate content from the archives.

EXTRACTING INFORMATION FROM TWITTER

To create incrementally enriched Web archives which allow access to all sorts of social media content in a structured and semantically meaningful way, we need to extract relevant information from the tweets (which can point to related information on the Web). Semantic technologies have the potential to help people cope better with social media-induced information overload, by making use of content from the social Web that is relevant to a specified topic, event, or entity that researchers and archivists may be interested in. From this information, we can also identify opinions, and track opinion changes over time, both of which help gauge public interest.

The content-collection process which we describe in the remainder of this chapter, in the form of information-extraction methodologies and crawling techniques and strategies, is under continuous development within the ARCOMEM project. Extraction covers the initial identification and structured representation of knowledge about events and entities from previously unstructured material. This process faces issues arising from diversity in the nature and quality of Web content, in particular when considering social media and user-generated content, where further issues are posed by informal use of language. Since archiving has to consider the evolution of content and metadata over time, temporal and dynamic aspects are of special importance. We aim to extract relevant information from tweets in order to answer questions such as:

- How did people talk about the issue or event?
- How are opinions distributed in relation to demographic user data?
- Who are the most active Twitter users?
- Who are the opinion leaders?
- Where did they come from?
- What did they talk about?
- How has the public opinion on a key person evolved?

ENTITY AND EVENT EXTRACTION

Information extraction from tweets involves the use of natural language processing (NLP) techniques to extract events, entities, and other kinds of information from the (unstructured) text of the tweets. The extracted information can then be used for targeted Web crawling, allowing the crawling strategy to be gradually refined according to some specification of the entities and events. A further challenge is then to make appropriate use of these outcomes to create focussed archives. Recognising occurrences of named entities (such as persons, locations, etc.) within a text can be broken down into two main phases: ontology-based entity annotation (or candidate selection) and entity linking (also called reference disambiguation or entity resolution). This is useful so that the entities extracted can be linked together (co-referenced), even when they appear in different documents, and disambiguated when multiple meanings are possible. For example, the word *Paris* could refer to the entities *Paris, Texas*, or *Paris, France*, or even *Paris Hilton*: we want to ensure that each time it occurs in a tweet, we know which of these it is referring to. We can also then group together all tweets that talk about Paris, Texas separately from those which talk about Paris, France. Ontology-based entity annotation identifies all mentions in the text of classes and instances from an ontology (such as DBpedia.org). The entity-linking step then uses contextual information from the text, as well as knowledge from the ontology, to choose the correct entity, associated with a unique identifier (uniform resource identifier, or *URI* in Semantic Web speak) in the case of ambiguity.

There are many tools and methods for extracting information from text, using both rule-based and statistical techniques. The extraction techniques used in the ARCOMEM project are all developed in GATE (Cunningham, Maynard, Bontcheva, & Tablan, 2002), an architecture for language engineering which contains a number of components for language processing and text mining. The extraction task can be broken down into the following tasks:

- document preprocessing (document format analysis, content detection);
- linguistic preprocessing (language detection, separating the text into words and sentences, annotation with simple grammatical features such as part-of-speech categories (nouns, verbs, etc.), and dependencies between them such as subject and object);
- entity and event recognition (ontology-based lookup, annotation using specific sets of rules, extraction of important terms and phrases, entity linking).

Traditionally, named entities are of the types Person, Location, Organisation, Date, Time, and Money. However, in some cases, we also want to extract entities specific to the domain in question. For example, for tweets about music events (rock concerts and so on), we might want to extract entities such as band names; for political tweets we might want to specifically extract political parties as a subtype of Organisation. Similarly, event types may be dependent on the domain: for example, music festivals have events such as performances, and sub-events such as incidents that happen during a band's performance. Usually, these specific types of event and entity will be predetermined, but there are also possibilities for creating and extracting new types on the fly, using techniques such as clustering of similar examples. And finally, event extraction involves the recognition of domain-important happenings or situations within which entities are related to each other.

OPINION EXTRACTION

Extracted entities and events can be used to drive the extraction of opinions from tweets. It is not enough in this case to simply know whether a tweet is positive or negative in general, but to know what exactly it is positive or negative about. It is thus important to relate the opinion to a target (topic); for example, a tweet may be negative overall (e.g., sadness about the death of a famous person) but positive about the actual person. We therefore use the entities and events as possible targets to which the opinions are anchored. Opinions and sentiments are first gathered at the sentence and word level from text-based documents, based on the recognition of sentiment referring to the entities and events previously identified; more information on how to do this is given by Maynard, Bontcheva, and Rout (2012). Opinions can then be aggregated over wider elements such as whole documents or individual blog posts, and stored along with the individual sentiments.

ISSUES WITH ANALYSING SOCIAL MEDIA

The analysis of tweets is challenging for text-mining systems because of their informal use of language and style. Typically, tweets are rich in abbreviations, slang, domain-specific terms, and spelling and grammatical errors. NLP techniques are usually developed to deal with standard language, and therefore tend to produce lower-quality results on this kind of informal text. For example, shortened or misspelled words increase the variability in the forms for expressing a single concept. One solution to this is the normalisation of text before processing, but this is not possible here because we wish to preserve the content in its original form. For example, misspelled entities need to be recognised as such, but also to be connected with the correctly spelled versions of the same entity. The quality of the text affects not only the actual recognition of entities but also all the linguistic processing components, such as part-of-speech (POS) taggers and so on, mentioned in the previous section. Degraded performance on any of these components may have a negative effect on any other components which rely on these, because they are run in series, with each depending on the results of the next. So the higher up the chain the error, the worse the knock-on effect; in particular, errors in tokenisation and POS tagging can severely hamper the entity and opinion extraction. For preprocessing, we can adopt a number of techniques, such as retraining the components specifically on tweets; using techniques from SMS processing; adding lists of emoticons; recognising artefacts such as smileys, @mentions, and hashtags separately; replacing common abbreviations with their full words (e.g., tnx = thanks); and so on. We can also adopt backoff strategies for dealing with informal text, such as using more flexible grammar rules and additional use of co-reference techniques: see Maynard et al. (2012) for a description and discussion of these.

CONTEXTUALISED TWITTER ARCHIVING

The previous section has mentioned a number of interesting features that can be extracted from the content of Twitter posts and the social networks they exist in. We now explain how to leverage this extracted information in the construction of focussed, contextualised archives of Twitter and Web data. This is done by using these features to guide a Web crawler.

The first step is for a Web archivist to specify the *scope* of the archive, with the scope specification relying on information extracted from Twitter (entities, social context, etc.) in addition to more traditional URL-based features. Once this is done, the archiving process can be launched. In contrast to more traditional Web crawling approaches, archiving Twitter requires using the Web APIs provided by Twitter, rather than conventional Web page crawls. Feedback from information extracted is then used to guide the crawler. We do not stop at capturing Twitter data—it is also important to crawl the content of the Web context of Twitter posts, in particular the URLs that Twitter posts point to. Finally, the archive can be enriched with a more in-depth analysis of its content.

ARCHIVE SCOPE

Traditionally, Web archivists and crawl engineers, when they use an archival Web crawler such as Heritrix (Mohr, Kimpton, Stack, & Ranitovic, 2004) to archive a part of the Web, express the scope of the indented crawl as a *crawl specification*. This is a document specifying a set of seed URLs, from which the crawl should be started, and a description of in-scope URLs, based on a whitelist and blacklist of URL patterns (typically described by regular expressions) and file formats (described by patterns on file extensions or MIME types). Such a specification may, for example, express that, for a given crawl, only resources under the *.gov.uk* domain name hierarchy should be archived, and that only HTML content together with some associated files (scripts, stylesheets, images) should be retrieved, excluding other kinds of content such as videos and PDF documents.

When one moves from regular Web archiving to the archiving of Twitter and associated social web content, this kind of crawl specification becomes too limited to express the scope of the archival process. Instead, in addition to regular URL seeds and URL patterns, an archival specification should consist, on the one hand, of keywords and key phrases relevant to the archive scope (e.g., "barack obama", "U.S. Politics") serving as *seeds* to search for on Twitter, and on the other hand, of a description of which structured entities (e.g., *Barack Obama*) or social network features (e.g., "users from the United States that are opinion leaders") are relevant. Essentially, anything that can be detected by the information-extraction components mentioned above can be added as a filter. All such components come with a *score* (a number between 0 and 1) quantifying relevance to the scope of the archive; this scope is then used to prioritise the crawler.

CAPTURING API CONTENT

Like any other Web site, Twitter can be crawled using a regular archival Web crawler. However, this is not the most efficient way to capture Twitter data, and it is usually preferable to access Twitter using its rich HTTP Application Programming Interface (see https://dev.twitter.com/ for the documentation). Indeed, the regular Web interface, which makes heavy use of AJAX to present information (presenting only a list of 20 or so recent tweets by default, with more being loaded asynchronously as the user scrolls down the Web page) is more cumbersome to use for retrieving content. The API, which provides 200 tweets at a time (for a user's timeline) as structured records of information, offers a wealth of different querying methods, such as *search* to discover tweets containing keywords and key phrases, or *streaming* to get a continuously updated list of tweets on a given topic (see Chapter 5 by Gaffney and Puschmann, in this volume).

The Twitter API restricts the number of requests that can be performed per hour (the precise amount depends on the method used). A Twitter API archiving system needs to be aware of this policy limitation in order to automatically adapt its rate of crawl. For this purpose, we have developed a general Social Web API crawler tool, *API Blender* (Gouriten & Senellart, 2012), that eases the burden of developing API-specific capturing tools that manage authentication, adapt to policy limits, and even transform the specific schema of the information presented by different social networking platforms (Twitter, Google+, Facebook, etc.) into a common, unified schema.

GUIDING THE ARCHIVING

Information extraction and social network analysis components drive the Twitter capture, in conjunction with the archive specification: once tweets are captured (starting with a search from the key phrases), they are analysed as described above (for example, to extract named entities) and their relevance to the crawl is assessed by a *prioritisation* module that decides whether to explore more of their context (social data, retweets, Web links), and determines the ordering of further requests to the API. Archival guidance can also come in a more indirect manner: once a capture has been made and the archive enriched and annotated (see Archive Enrichment below), the archive specification can be refined, and another capture can be launched, to focus more on those parts of the Twitter social network that were judged important.

CRAWLING THE CONTEXT

Building an archive of Social Web content is more than just building an archive of tweets: it is also critical to crawl the Web context of these tweets, in the form of the Web resources referenced in tweets, and possibly neighbouring pages thereof. The Twitter API capturing system thus needs to extract all hyperlinks found in tweets, if deemed relevant to the archive specification, and hand them over to a regular Web crawler. This regular Web crawler, in turn, uses these URLs as seeds, and crawls the corresponding Web content, also applying the scoring and filtering criteria defined by the crawl specification.

Conversely, once the Web crawler encounters the URL of a Twitter user, it makes sense for it to delegate the capture of the corresponding content to API Blender, by transforming the URL into the corresponding API method call.

One technical problem is raised by the common use of URL shorteners (HTTP redirection services that replace a long URL with a shorter one such as http://bit.ly/dG6yFL). Indeed, it is often the case that URLs make use of a chain of shorteners: they use a generic URL shortener in addition to Twitter's own, mandatory shortener http://t.co/. The use of these URL shorteners makes it harder for the information-extraction components to estimate the relevance of a given link to an archive specification, since nothing in the URL indicates its content. The URL must therefore be resolved before assessing it.

ARCHIVE ENRICHMENT

The result of the focussed archiving guided by information extraction described above is an archive that can be further enriched with metadata on attributes such as entities and opinions. (The extraction of these attributes has been described above.) Archive enrichment is an important aspect of social media preservation, because it enhances the quality and usefulness of the content. It enables different perspectives on the data to be encoded and searched, since archive users can search not just by the content of texts but also by the metadata attributes assigned to them. For example, they can investigate particular opinions about certain entities, look for changes to these over time, and perform other complex, information filtering processes, thus inferring new knowledge from the captured, enriched, and contextualised Web content.

ARCHIVE USAGE

Given the fast growth of social media exploitation, it is to be expected that the use of social media in general, and of Twitter content in particular, will rapidly extend to all areas of professional activity where organisations profit from gaining insight into the social repercussions of issues that are closely related to these organisations' interests. Social networks are a rich information source, whose structures can be exploited to acquire knowledge about facts and opinions (Dietze et al., 2012) as well as social connections and interactions (Agichtein, Castillo, Donato, Gionis, & Mishne, 2008). This is recognised by an increasing number of stakeholders in a wide variety of application domains for Web archiving. In addition to their traditional sources (news agencies, PR material, or library content), professionals such as archivists and journalists want support for selecting and archiving relevant, user-generated content from tweets, in order to preserve this ephemeral content, and to enable the retrieval of relevant, tweet-derived source material. Stakeholders with an interest in the appraisal of their products in social media environments, such as media organisations and political actors, will be able to mine and follow societal feedback for short-term purposes. Beyond this, although short-term storage is required when immediate use is important, additional storage strategies are necessary for longer-term preservation.

LONG-TERM PRESERVATION

The long-term usage of Twitter archives, and of Web archives more generally, raises a range of issues. Archived content should be kept accessible and usable well into the future. Also, access to the archive, as well as to a contextual understanding of the content at the time of publication, should be supported.

First challenges arise when the technological development concerns the accessibility and interpretability of the content (CCSDS, 2012). In the worst case, there are no tools available to present the content of the archive in an intelligible form. To avoid this situation, the usage of standards that are supported by a wide range of tools and maintained over time is an obvious necessity. Capturing Twitter results in a substantial number of JSON files (IETF, 2006). JSON (JavaScript Object Notation) is a lightweight, text-based, language-independent, data interchange format standardised by the Internet Engineering Task Force (IETF). The documentation is publicly available and widespread. Therefore, the semantics of the information items within a JSON file are well-documented for future usage.

However, preserving the JSON files alone is not sufficient. There are also many different forms of technical and descriptive meta-information that need to be preserved. The Web ARChive file format (WARC) has been standardised by the International Organisation for Standardisation (ISO, 2009). This choice of format for the long-term preservation of Web resources of all kinds is an accepted standard in the archive community. WARC archives aggregate multiple resources into a single file. Besides the content, it also stores related meta-information.

To ensure the technical accessibility and usability of the archive content is one step for the long-term usage of Twitter archives. As discussed above, individual tweets are limited in their length and contain very little information, which complicates the intelligibility of the content at a later time. A Twitter message such as "The new #ipod is cool http://bit.ly/NWou" will hardly be understandable in 50 years without additional knowledge. It is impossible to predict whether future users of the archive will have information on what an iPod was in 2012. While traditional materials, such as papers or books, often provide enough contextual information to be intelligible, this is rarely the case for user-generated content on the Social Web. Therefore, as much context information as possible—like descriptions of major entities and concepts (such as the concept of a portable media player, in the iPod example)—should be kept together with the tweet. This will not guarantee full intelligibility, but it is an important step in that direction.

We have outlined above how identified entities and concepts could be connected to the linked data cloud, for example by referencing DBpedia. When searching across long-term archives, different instances of a concept such as "portable media player" might occur: for example, Walkman, Watchman, Discman, MP3 Player, iPod. URIs linking to DBpedia or Wikipedia as references to an entity can help to identify information objects with similar semantics. The benefit for the reader is that they can access contextual information in terms of related documents as well as the description of an entity. This ensures the long-term semantic interpretability of the content.

CONCLUSION

Capturing tweets together with their context (if a context link or other context information is provided) allows for a better understanding of individual messages and groups of tweets. Therefore, a comprehensive mechanism for the archiving of Twitter content must consist of capturing API content as well as regular Web crawling, in order to collect both types of information. The enrichment of the captured data enhances the subsequent access and usage of the archives which this mechanism creates.

REFERENCES

Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008). Finding high quality content in social media, with an application to community-based question answering. In *Proceedings of Web Search and Data Mining (WSDM)* (pp.183–194). Stanford, CA: ACM Press.

ARCOMEM. (2012). ARchive Communities' MEMories (ARCOMEM). Retrieved from http://www.arcomem.eu/

CCSDS. (2012). Reference model for an Open Archival Information System (OAIS). Magenta Book. Issue 2 June 2012. Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)* (pp. 168–175) Philadelphia, PA.

Dietze, S., Maynard, D., Demidova, E., Risse, T., Peters, W., Doka, K., & Stavrakas, Y. (2012). Entity extraction and consolidation for social web content preservation. In *Proceedings of 2nd International Workshop on Semantic Digital Archives (SDA)* (pp. 18–29), Pafos, Cyprus.

Gouriten, G., & Senellart, P. (2012). API Blender: A uniform interface to social platform APIs. In *Proceedings of the 21st World Wide Web Conference (WWW 2012), Developer Track*. Retrieved from http://www2012.wwwconference.org/proceedings/nocompanion/DevTrack_039.pdf

IETF. (2006). The application/json media type for JavaScript Object Notation (JSON). Retrieved from http://www.ietf.org/rfc/rfc4627.txt

ISO. (2009). Information and documentation— The WARC file format (ISO/DIS 28500). Retrieved from http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44717

Maynard, D., Bontcheva, K., & Rout, D. (2012). Challenges in developing opinion mining tools for social media. In *Proceedings of @NLP can u tag #usergeneratedcontent?! Workshop at*

LREC 2012, May 2012, Istanbul, Turkey. Retrieved from http://gate.ac.uk/sale/lrec2012/ugc-workshop/opinion-mining-extended.pdf

Middleton, M. (2012). *Defining Web archive scope*. Retrieved from http://web.hanzoarchives.com/bid/90416/Defining-web-archive-scope

Mohr, G., Kimpton, M., Stack, M., & Ranitovic, I. (2004). Introduction to Heritrix, an archival quality Web crawler. *Proceedings of the 4th International Web Archiving Workshop (IWAW 2004)*. Retrieved from http://www.iwaw.net/04/Mohr.pdf

Mourtada, R., & Salem, F. (2011). Civil movements: The impact of Facebook and Twitter. *Arab Social Media Report*, *1*(2). Retrieved from http://www.dsg.ae/En/Publication/Pdf_En/DSG_Arab_Social_Media_Report_No_2.pdf

Speak2Tweet. (2012). https://twitter.com/speak2tweet

Twitter first off the mark with Hudson plane crash coverage. (2009). Retrieved from http://www.editorsWeblog.org/2009/01/19/twitter-first-off-the-mark-with-hudson-plane-crash-coverage