Twitter Integration and Content Moderation in GDACSmobile

Daniel Link
Chair of Information Systems and Logistics,
European Research Center for Information Systems (ERCIS)
University of Münster, Germany
daniel.link@ercis.de

Bernd Hellingrath
Chair of Information Systems and Logistics,
European Research Center for Information Systems (ERCIS)
University of Münster, Germany
bernd.hellingrath@ercis.de

Tom De Groeve
Joint Research Centre of the European Commission
tom.de-groeve@jrc.it

ABSTRACT

Recent years have shown that mobile devices and Twitter can play a significant role in providing real-time data from disaster-affected areas to disaster managers. Against this background we present a workflow for Twitter integration into a disaster management information system, and a concept for content moderation that can increase the quality of disseminated information.

Keywords

Coordination, GDACS, GDACSmobile, needs assessment, social media, Twitter, content moderation

INTRODUCTION

Social media currently is one of the most discussed sources for real-time data for disaster management. Among many social media sources, Twitter became a frontline candidate because of its status of being a "what's-happening-right-now"-tool (Bifet & Frank, 2010). The platform which allows its users to share 140 character long messages contains timely disaster information as shown in the works of (Hermida, 2010; Hughes & Palen, 2009; Jansen *et al.*, 2009; Shklovski *et al.*, 2010) and many others. This is emphasized by the use of Twitter data in several software applications that have been successfully employed in the field of disaster management, such as Ushahidi and Twitcident (Morrow *et al.*, 2011; Ushahidi, 2013; Twitcident, 2012). The reasons for sharing disaster relevant information within social media were found in organizational and intrinsic people behaviour (Hughes & Palen, 2009; Starbird & Stamberger, 2010), striving for reducing uncertainty caused by disasters (Shklovski *et al.*, 2010) and in the general willingness of people to use ICT during disasters (Hughes & Palen, 2009; Sheetz *et al.*, 2010). All those aspects make the integration of Twitter into the existing information landscape of disaster management particularly interesting.

Our research group from the University of Münster and the Joint Research Centre of the European Commission aimed at developing a solution that disaster relief professionals and the affected population can use to acquire and disseminate real-time data within the first four weeks after a major sudden-onset disaster.[1] The resulting solution "GDACSmobile" enables disaster management professionals and the affected population to share their observations from the affected area as "situation reports", both via the GDACSmobile client application for mobile devices, and via Twitter. Via the client application professionals and population can also receive

[1] As our main guidance for development, we relied on the Three Layers Design Guideline, because it covers the complete design process, and focuses on mobile applications (Ayob *et al.*, 2009). By following this approach, we applied the Human-Centred Design (HCD) guideline as an overall framework, and the golden rules of Shneiderman that have been adapted to the design of mobile applications (Gong & Tarasewich, 2004; Shneiderman & Plaisant, 2004). To enhance usability, we additionally considered a design guideline for context-aware mobile applications (Häkkilä & Mäntyjärvi, 2006).

situation reports that provide valuable information to their decision-making.[2] To ensure the quality of the information that is disseminated through the client application, the solution employs a concept to moderate incoming reports.

In this paper we present the general workflow that is used in the solution as well as its key objects. Based on this, we focus on how we integrated Twitter, and how the content moderation works. Eventually we draw a conclusion and give an outlook on possible future work.

GENERAL WORKFLOW

The GDACSmobile target groups, i.e. disaster management professionals ("authorized users") and affected population ("public users"), share their observations from the disaster-affected area. They do this by sending reports to the provider's server via the client application or via Twitter. How the client application asks the users to describe an observation is determined by the report structure that is configured on the server.

The provider uses the server application to receive user reports and to moderate them (see "Moderation Concept", p. 70). By receiving reports the provider gains an overview of the situation, and may choose to re-configure the report structure on the server. For instance, the provider may add a specific question about the health services that the affected population can access. When a client next accesses the server, it receives the re-configured report structure. The client then also receives new reports that are passed through moderation. Receiving both the current report structure and current situation reports closes the assessment cycle. With the newly gained information, users can better react to the situation they are in, e.g. by using a reported local transport capacity for delivering aid.

KEY OBJECTS

The key objects within the solution are: report, category, user, alert, and mission.

- A "report" describes an observation made by a user. It belongs to exactly one instance of each of the objects: category, mission and user, which provide meta-information for categorization, filtering and search.
- To create a report using the client application, the user has to select a predefined "category", which specifies the context of the report. Categories can be organized hierarchically, i.e. root categories can contain sub-categories. Once a user has selected the appropriate category for his observation, the application provides further guidance by asking specific questions that match the category. This is done by assigning a category to a report template that contains assessment questions.[3] For instance, a user who wants to share the information that diarrheal disease is spreading in his area, the user may create a report in the water and sanitation category and in the diarrheal disease sub-category. There the user is asked to specify the number of people affected.
- A "user" can submit reports. Users are either authorized professionals or anonymous public users. Authorized users are disaster management professionals who are registered within the system, e.g. by linking GDACSmobile and VirtualOSOCC[4] accounts. Public users do not need to be registered, but can still be tracked through an identifier. The identifier refers either to their mobile device or to their Twitter account.
- An "alert" identifies a certain disaster. Alerts are imported each hour from the GDACS alert newsfeed. For each new alert a public mission is automatically created.

A "mission" groups reports about one particular alert. There can be multiple missions per alert.[5]

[2] The client application can be used without any access to the Internet, e.g. for recording observations. An Internet connection is only needed to receive automatic updates, to load new map tiles, and to send reports to the server that were created since the Internet connection became unavailable.

[3] The templates are flexible and can be re-configured on the server to reflect the unique and dynamic characteristics of disasters. Following the example, when the provider receives multiple observations about diarrheal disease from a specific area, it might be useful to ask where from people get their water.

[4] The United Nations Office for the Coordination of Humanitarian Affairs (OCHA) provides information required by the international community to support humanitarian response. OCHA does this also via the web platform VirtualOSOCC (Virtual On-Site Operations Coordination Centre), which is accessible by disaster managers through the GDACS website (European Commission, 2013).

[5] By default there is a public mission for each alert that any user can select to share observations. Apart from the public mission, there can also be multiple private missions for each alert. A private mission resembles a

TWITTER INTEGRATION

To make good use of the real-time data that Twitter can provide, it is necessary to understand why emergency response organizations mainly use Twitter for dissemination only, instead of using it to coordinate relief efforts (Tapia *et al.*, 2011). Perhaps the most striking reason lies in the uncertainty connected to Twitter data (Coyle & Meier, 2009). Virtually every bit of data coming from Twitter is potentially unreliable and has to be verified, e.g. because it could have been placed for the purposes of misinformation and propaganda (Burns, 2010; Coyle & Meier, 2009). This is a problem, since departments that are in charge of immediate disaster response have only little time for verifying uncertain information before they must act (Tapia *et al.*, 2011). Disaster managers can also allocate only few resources to extensively try out new technologies until the technologies prove to be effective and reliable (Tapia *et al.*, 2011).

Considering this background, we decided to employ bounded microblogging as the primary approach for Twitter integration, as described in the following, mainly according to (Tapia *et al.*, 2011). Bounded microblogging means that only the messages of a circle of selected users are considered. The benefits of bounded microblogging lie in the trustworthiness of the collected information, and in the few resources required for implementing the approach. It is also a promising starting point to introduce the medium Twitter to emergency response organizations for other purposes than public relations. An important drawback is that data from the majority of Twitter users is ignored, leaving only the small circle of selected users to be considered, and hence the data does not scale. This drawback is only a little weakened by the fact that 50 per cent of the messages on Twitter are sent by only 0.05 per cent of the total users (Wu *et al.*, 2011).

To also make use of the information that non-selected Twitter users contribute, we developed the workflow that is outlined in Figure 1.

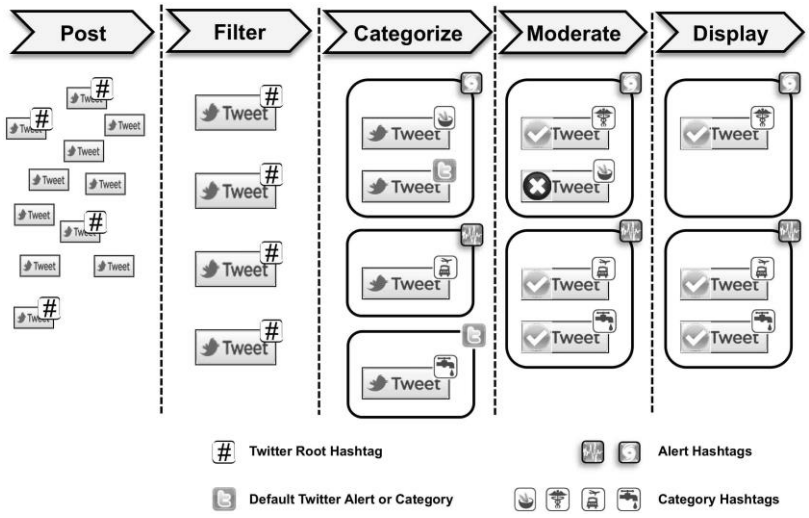

Figure 1. Twitter Workflow

In a pre-stage to this workflow, Twitter users have to be informed about the hashtags they should use to share their observations. This includes a mandatory root hashtag that makes a Twitter message (tweet) visible to the system, as well as alert and category hashtags that aid processing. Then, in the first step of the workflow, the server receives all tweets that include the root hashtag, e.g. "#gdacsm". In the second step, filters are applied, e.g. only tweets are regarded that provide the alert hashtag "#eqsf122" for "Earthquake, San Francisco, 2012, no. 2" and contain geographic location coordinates. In the third step, the system transforms each tweet into a report, assigns it to the appropriate alert, and categorises it using the available category hashtags. If no category hashtags are available it assigns the tweet to the default Twitter category. From this moment on, tweets are treated like reports from public users, which have to be reviewed by a human moderator before they are displayed to GDACSmobile users. After the content of a Twitter message is validated and perhaps corrected, e.g. by adjusting its geographic location information, the report status is revised accordingly. The various kinds of report statuses are explained below.

---

password-protected space that can be used by a group of users to share their observations about the particular alert. Every report within that private space is automatically shared with its members, and requires moderation to be shared with users who do not have the password to the mission. For instance, an emergency response organization can use a private mission to first collect all observations made by one of their assessment teams, and process the information before releasing it to other organizations and the public.

MODERATION CONCEPT

The employed moderation concept ensures the quality of the information that is disseminated through the GDACSmobile client application. In the concept, every report assumes a status, as shown in Table 1.

Status	Default status for reports		Report visible on client	
	Authorized User	Public User	Authorized User	Public User
Accepted	X	-	X	-
Public Accepted	-	-	X	X
Not Evaluated	-	X	-	-
Rejected	-	-	-	-

Table 1. Possible Report Statuses

Reports from public users (incl. tweets) take on the status "not evaluated" by default, meaning that they have to undergo moderation before they can become visible to other users. In contrast, reports that have been submitted by authorized users are "accepted" by default. This is because we assume the professional background of authorized users to indicate their trustworthiness. When a report assumes the status "accepted" it becomes visible to authorized users, and when it is "public accepted" it becomes visible to both authorized and public users. Users cannot view "rejected" reports, as it should be in the case of false or misleading information.

Moderators log on to the server using a web browser, and view the report list. A moderator can apply various filters to the list of reports. Filter criteria are alert (disaster), category, data source, and report status. This can be the basis of flexible workflows and the distribution of labour among moderators, e.g. by one moderator categorizing tweets, and a second moderator checking the validity and "accepting" reports within the "infrastructure" category and the "roads and bridges" sub-category. When moderators view individual reports, they see all attributes of the report, e.g. its location (coordinates and map) and the fields that were defined in the report template. Moderators can alter the report if needed, add a comment, and choose to change the report status. Upon status change, the next report in line is automatically loaded to speed up review. It is also possible for moderators to filter "accepted" or "public accepted" reports and sort the reports by time of submission (e.g. oldest first) for the sake of checking the validity of old reports.

By adding moderation to the bounded microblogging approach, it is possible to apply "Microblogging as Ambient or Context", similarly to how it is described by (Tapia *et al.*, 2011). That is, the moderator may choose to enrich the existing information from authorized users with the contextual information provided by public users, because it adds ambient or contextual information to their observations. The moderator can also choose to verify the received information, e.g. by asking a Twitter user who is located close by.

CONCLUSION AND OUTLOOK

In this paper we motivated the inclusion of Twitter into the GDACSmobile solution by highlighting its relevance for disaster management. We presented the solution's key objects, and on this foundation we described how Twitter and content moderation enable the combination of bounded microblogging and Twitter as ambient or context.

Initial tests of the GDACSmobile solution have been performed in the vicinity of the development area on multiple device types. The test run was focused on the typical operational lifecycle of GDACSmobile, i.e. report submission either by Twitter or GDACSmobile client as well as content moderation on the server. As the test was meant to evaluate the real life behavior with a high number of reports coming in, a long running server with about 4,000 reports was used for all test runs. It has to be noted that a field test in a developed country with (potentially) sophisticated mobile networks does not represent the full situation found in a real life disaster. To further test the solution, we intend to deploy it within a larger scale simulation. The simulation should involve a higher number of testers, including both users without disaster experience and disaster management professionals.

Furthermore, the effectiveness and value of tweets within the GDACSmobile solution can be evaluated in the future. Based on the evaluation results, it might be promising to integrate further social media platforms. For instance, Flickr could be used to retrieve relevant pictures from the affected area.

The initial categories for the solution were the result of the development of a minimal situation report structure for disaster management, which is not in the focus of this paper. We intend to refine the structure in general, and extend it to reflect the information requirements of humanitarian logistics in particular.

ACKNOWLEDGMENTS

We want to thank the team of students at the Research Group on Information Systems and Supply Chain Management for their contribution to the project: Carsten Bubbich, Friedrich Chasin, Sven Kronimus, Stefan Laube, Philipp Saalmann, and Martin Vanauer. For sharing his views in many discussions during the GDACSmobile development, we also want to thank Minu Limbu, who has years of experience in the field of disaster management in general, and information management in particular.

REFERENCES

1. Ayob, N., Hussin, A. and Dahlan, H. (2009) Three Layers Design Guideline for Mobile Application, *International Conference on Information Management and Engineering*, Kuala Lumpur, Malaysia, 428.
2. Bifet, A. and Frank, E. (2010) Sentiment Knowledge Discovery in Twitter Streaming Data, *Proceedings of the Discovery science: 13th International Conference, DS 2010*, Canberra, Australia, 1-15.
3. Burns, A. (2010) Oblique strategies for ambient journalism, *MC Journal*, 13, 2.
4. Coyle, D. and Meier, P. (2009) *New technologies in emergencies and conflicts: The role of information and social networks*, United Nations Foundation, Washington, D.C.
5. European Commission (2013) GDACS Website 2.0 [WWW document], URL http://www.gdacs.org/, accessed 15 February 2013.
6. Gong, J. and Tarasewich, P. (2004) Guidelines for Handheld Mobile Device Interface, *Proceedings of the 2004 DSI Annual Meeting*, Boston, MA.
7. Häkkilä, J. and Mäntyjärvi, J. (2006) Developing design guidelines for context-aware mobile applications, *3rd international conference on Mobile technology, applications & systems*, Bangkok, Thailand, 1-7.
8. Hermida, A. (2010) Twittering the news: The emergence of ambient journalism, *Journalism Practice*, 4**,** 3, 297-308.
9. Hughes, A. L. and Palen, L. (2009) Twitter adoption and use in mass convergence and emergency events, *International Journal of Emergency Management*, 6**,** 3/4, 248.
10. Jansen, B. J., Zhang, M., Sobel, K. and Chowdury, A. (2009) Twitter power: Tweets as electronic word of mouth, *Journal of the American Society for Information Science and Technology*, 60, 11.
11. Morrow, N., Mock, N., Papendieck, A. and Kocmich, N. (2011) Independent Evaluation of the Ushahidi Haiti Project [WWW document], URL http://www.alnap.org/pool/files/1282.pdf, accessed 28 November 2012.
12. Sheetz, S. D., Kavanaugh, A., Quek, F., Kim, B. J., and Lu and S. C. (2010) Expectation of connectedness and cell phone use in crisis, *International Journal Of Emergency Management*, 7**,** 2, 124.
13. Shklovski, I., Burke, M., Kiesler, S. and Kraut, R. (2010) Technology Adoption and Use in the Aftermath of Hurricane Katrina in New Orleans, *American Behavioral Scientist*, 53, 8, 1228-1246.
14. Shneiderman, B. and Plaisant, C. (2004) Designing the user interface: Strategies for effective human-computer interaction, Addison-Wesley.
15. Starbird, K. and Stamberger, J. (2010) Tweak the Tweet: Leveraging Microblogging Proliferation with a Prescriptive Syntax to Support Citizen Reporting, *Management*, 1-5.
16. Tapia, A. H., Bajpai, K., Jansen, B. J. and Yen, J. (2011) Seeking the Trustworthy Tweet : Can Microblogged Data Fit the Information Needs of Disaster Response and Humanitarian Relief Organizations, *Proceedings of the 8th International ISCRAM Conference*, Lisbon, Portugal, 1-10.
17. Twitcident (2012) Twitcident - Helping Emergency Services [WWW document], URL http://twitcident.com, accessed 15 February 2013.
18. Ushahidi (2013) Ushahidi [WWW document], URL http://ushahidi.com/, accessed 15 February 2013.
19. Wu, S., Hofman, J. M., Watts, D. J. and Mason, W. A. (2011) Who Says What to Whom on Twitter, *Proceedings of the 20th international conference on World wide web*, New York, NY, 705-714.