Provenance and Social Machines

Milan Markovic, Peter Edwards, David Corsar and Jeff Z. Pan
Computing Science & dot.rural Digital Economy Hub
University of Aberdeen
Aberdeen, AB24 5UA
{m.markovic, p.edwards, dcorsar, j.z.pan}@abdn.ac.uk

ABSTRACT

Social machines that outsource tasks to the crowd often have to address issues associated with the quality of contributions. In this paper we discuss a solution based on the maintenance and use of a provenance record.

Categories and Subject Descriptors

H.1.2 [**Information systems**]: Models and Principles

General Terms

Theory, Management, Design

Keywords

Social Machine, Linked Data, Crowdsourcing, Data Streams, Provenance

1. INTRODUCTION

Berners-Lee and Fischetti in "Weaving the Web" [2] describe the concept of *social machines* as: "processes in which the people do the creative work and the machine does the administration...". Such machines thus blend the capabilities of both humans and machines to perform tasks that machines alone would be unable to perform. Examples of social machines can be found today in the form of crowdsourcing systems, which outsource jobs to a large group of people via an open call [3]. The openness of such systems requires them to be able to tackle problems such as imperfect data [7], and error generating participants [4]. Typical solutions are based on the use of the crowd to perform some form of validation (e.g. through the use of voting or rating systems). However, in situations where the number of potential participants is small (e.g. rural areas, scarce expertise) such solutions might prove problematic. If more information was known about the validator (e.g. level of expertise), then a system could operate with limited number of such participants. We argue that such validation (e.g. voting or rating systems) can be improved if combined with automatic data evaluation processes (e.g. data quality, or trustworthiness/reputation of participants). However, performing such evaluations requires additional contextual information, which is often unavailable (e.g. how the data was created, or who created it). We consider linked data principles [1] - a set of principles for consuming and publishing machine-readable data on the web, as an underpinning infrastructure facilitating the acquisition of required contextual information. Previous research has identified provenance as essential for supporting information discovery and assessments such as reliability and quality [9]. We are therefore exploring the use of provenance to provide additional contextual information required by automatic data evaluation processes within social machines.

Digital Futures '12 Aberdeen, Scotland UK

2. SAMPLE APPLICATION

Monitoring of travel disruptions in rural areas is often difficult and poses several challenges (e.g. how to obtain information from the site of an incident). A crowdsourcing application able to gather, manage, and assess disruption reports would provide an obvious solution. Figure 1 presents an outline architecture for such a system built around the linked data principles.

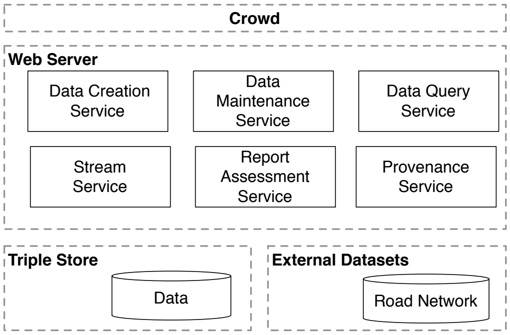


Figure 1: System architecture.

The system allows participants to report travel disruption events (e.g. an accident on a particular route) using their mobile device. In addition, they are able to perform tasks such as the creation of links between disruption reports, or maintenance tasks such as validation of data provided by other participants. By linking here we mean the identification of relationships between disruption reports. Figure 2 illustrates a scenario in which a number of reports about queueing traffic have been contributed to the system following a report about a car accident. To link/associate separate events and then identify all the related effects of the car accident (such as queues) would be a difficult task for a machine. However, crowd members with local knowledge of the road

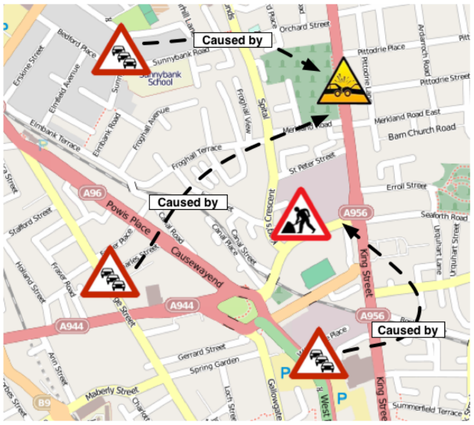


Figure 2: A map illustrates a number of disruption reports (e.g. queues, accident, road works) and the causal relationship between them.

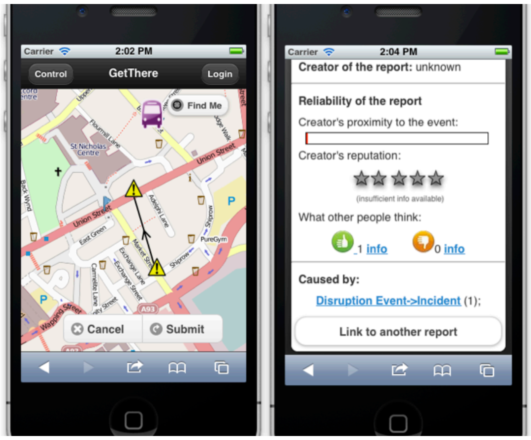
network and local commuting habits would be ideally suited to this task. Creating, linking, and maintaining the disruption reports alone does not provide important contextual detail such as who created it, who performed a maintenance operation, or when and how it was performed - all of which are useful when assessing the credibility of participants and the data they contribute. We argue that a provenance record is required to provide this context, by capturing information about participants and their activities. We adopt the W3C Provenance Working Group[1] definition of provenance as "a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing"[8].

A prototype of this system (Figure 3) is being developed as part of the Informed Rural Passenger Project[2], with the goal of creating a transport information ecosystem. Within this system issues such as data provenance, reliability of passenger-sourced information, and travel behaviour change are being explored. The prototype client application collects information from the crowd and communicates results back to users. It is optimised for use on touch screen mobile devices and supports the creation of disruption reports, validation reports, and reports about relationships (links) between disruption reports. It also provides a means to visualise reports and their associated links. The visualisation of disruption reports includes additional contextual information such as: reputation of the report creator, his proximity to the event at the time the report was created, and validation reports (opinions) of other people about this particular event. The system further consists of a server-side framework, which was built as a set of RESTful web services. The prototype is further described in [6].

3. APPROACH

3.1 Managing Crowdsourced Data

A disruption event report (as described in section 2) is likely to trigger a stream of data relating to this event, such as other disruption reports, or validation reports. It is therefore entirely natural to represent these data as a stream of elements, with participants contributing to a stream about a particular event (e.g. an incident on route A90). A system utilising the crowd to manage travel disruption would thus need to be built around a set of such streams. The participants contributing to these streams can undertake different roles such as data creation and data maintenance [5]. The participants create new data either within existing datasets or as part of new datasets (e.g. creating travel disruption reports). A linked data representation of these objects (e.g. disruption reports) is then generated. The participants can also be used to define new links between datasets, either as alignments (i.e. defining equivalent concepts) or as new relationships between previously unlinked concepts. Two data maintenance tasks that can be performed by the participants are validation and editing. Here validation involves the participants (validators) evaluating data and annotating them according to some quality or correctness vocabulary. Editing is then the process of revising data that has been previously annotated as being of poor quality.



Creator of the report: unknown

Reliability of the report

Creator's proximity to the event:

Creator's reputation:

(insufficient info available)

What other people think:

1 info 0 info

Caused by:

Disruption Event->Incident (1);

Link to another report

Figure 3: A prototype mobile client for reporting travel disruption. Left: A user links two related reports. Right: An overview of additional information associated with a disruption report.

Capturing the provenance of a stream object (e.g. the disruption report that initiated the stream) and the provenance of stream elements (e.g. who created a specific data element, or created a link between elements) would provide additional context to support reasoning about the quality of the data on the stream.

3.2 Modelling Provenance

The W3C Provenance Working Group[3] is defining a new standard model for the interchange of provenance information. The main components of the model are: entities (a thing with some fixed aspects associated with it), activities (something that occurs over a period of time and acts upon or with entities), agents (something that is responsible for an activity taking place), and relationships between these elements. In our work we distinguish between two types of provenance that can be modelled within a

[1] http://www.w3.org/2011/prov/wiki/Main_Page
[2] www.dotrural.ac.uk/irp

[3] http://www.w3.org/2011/prov/wiki/Main_Page

social machine: data provenance and stream provenance.

Data provenance is generated in response to a number of events: when data is created; when data arrives from the participants; or when links between contributions are created. The data provenance record then contains information such as the agent that created the data, the activities involved in creating the data (e.g. acquiring the agent's location, uploading from the client application, and subsequent processing by web services), and the entities used/generated by these activities.

Stream provenance is generated in response to: creation of a new stream; closing a stream; and data being added to a stream. The stream provenance record then contains information such as the activities that triggered the creation or closure of a stream, the activities that added elements to a stream, and the entities used by those activities (e.g. the data that was received from participants).

4. DISCUSSION

Social machines in the form of crowdsourcing systems exist in many domains such as transport[4] and healthcare[5]. We have described how the use of linked data, streams and provenance within such systems can be used to address issues such as the quality of contributions. The provenance record provides an audit trail that can support, for example the discovery of participants who generate reports that are frequently edited by validators. This in turn may form part of a reliability assessment of those crowd members, and assessment of the quality of their outputs. These types of analysis can also aid processes such as selecting a workforce for future applications, or monitoring/evaluating crowd performance. However, there are several issues associated with the use of provenance in this way: it may be possible to create only very limited provenance graphs; ensuring links within the graph are correctly generated; referencing items not published as linked data; and referencing triples deleted as part of an edit performed during maintenance.

For the purposes of our research we have identified the flowing research questions: *How can we capture the requirements for provenance within social machines built on linked data principles?*; *Can existing provenance models capture the non-stream aspects of such machines?*; *Within such machines, how can provenance of elements within a data stream and the data stream itself be represented?*; *What are the practical challenges of embedding provenance in social machines built on linked data principles?*.

At the current stage of our research we are focusing on answering the first of these questions through the process of: identifying various use cases for provenance within social machines; summarising the requirements identified from each scenario and producing a generic list of requirements for handling provenance within social machines which is essential for the process of answering the second research question.We plan to evaluate the list of requirements against the capabilities of existing provenance models. Evaluation of the results will suggest if new models for handling provenance in the context of social computation are required. Additional potential contributions include new provenance-driven methods for reasoning about crowd members. We plan to develop new reasoning methods for assessing the trust and reputation of the participants within social machines, which make use of the provenance information. In particular, we will focus on scenarios where the provenance information could support processes in which the crowd also becomes the validator of the data it generates (e.g. how do we trust other people that validated an incident report). Crowdsourcing systems are often required to process large data volumes which arrive into the system in the form of a stream (e.g. different people contributing at different times). We are therefore exploring the capability of current provenance models to handle the stream provenance as discussed earlier in this paper. This might lead to the extensions to current provenance model representations to support data streams.

5. ACKNOWLEDGMENTS

The research described here is supported by the award made by the RCUK Digital Economy programme to the dot.rural Hub; award: EP/G066051/1.

References

[1] T. Berners-Lee. Linked data. *http://www.w3.org/DesignIssues/LinkedData.html*, accessed:10/03/2012.

[2] T. Berners-Lee and M. Fischetti. *Weaving the Web: The original design and ultimate destiny of the World Wide Web*. Harper Collins, NY, 1999.

[3] J. Howe. The rise of crowdsourcing. *Wired Magazine*, 14(6), June 2004.

[4] J. Le, A. Edmonds, V. Hester, and L. Biewald. Ensuring quality in crowdsourced search relevance evaluation. In *SIGIR Workshop on Crowdsourcing for Search Evaluation*, pages 17–20, 2010.

[5] M. Markovic, P. Edwards, D. Corsar, and J. Z. Pan. The crowd and the web of linked data: A provenance perspective. In *AAAI Spring Symposium "Wisdom of the Crowd" Technical Report SS-12-06*, pages 50–51. AAAI Press, 2012.

[6] M. Markovic, P. Edwards, D. Corsar, and J. Z. Pan. Managing the provenance of crowdsourced disruption reports. In *Proceedings of the 4th International Provenance and Annotation Workshop-IPAW 2012*, 2012.

[7] P. Marsden. Crowdsourcing. *Contagious Magazine*, pages 24–28, 2009.

[8] L. Moreau, P. Missier, K. Belhajjame, R. B'Far, J. Cheney, S. Cooppens, S. Cresswell, Y. Gil, P. Groth, G. Klyne, T. Lebo, J. McCusker, S. Miles, J. Myers, S. Sahoo, and C. Times. Prov-dm: The prov data model. W3C Working Draft 24 July 2012, http://www.w3.org/TR/prov-dm/, July 2012.

[9] Y. L. Simmhan, B. Plale, and D. Gannon. A survey of data provenance in e-science. *SIGMOD Rec.*, 34:31–36, September 2005.

[4] http://highwire-dtc.com/ourtravel/?page id=195

[5] http://www.sickweather.com/