Following Tweets Around

Informetric methodology for the Twittersphere

David Gunnarsson Lorentzen

Academic dissertation for the Degree of Doctor of Philosophy in Library and Information Science at the University of Borås to be publicly defended on Monday 3 October at 13.00 in lecture room C203, the University of Borås, Allégatan 1, Borås

The Swedish School of Library and Information Science

The University of Borås

Title: Following tweets around: Informetric methodology for the Twittersphere

Language: English, with a summary in Swedish

Available: http://urn.kb.se/resolve?urn=urn:nbn:se:hb:diva-9339
ISBN (printed version) 978-91-981653-0-2
ISBN (digital version) 978-91-981653-1-9

The purpose of this thesis is to critically discuss methods to collect and analyse data related to the interaction and content on the social platform Twitter. The thesis contains examples of how networked communication can be studied on Twitter, based on the affordances of the platform considering interaction with interfaces and other users. The foundational problem is that social science Twitter research has been based on easily accessible data without introducing or discussing criteria for collecting appropriate samples for a given research task.

The thesis builds on one literature review and four studies of political Twitter communication. The analyses are based on a view of the Twitter platform as a non-neutral filtering gatekeeper. On the one hand, Twitter treats content and users asymmetrically, by emphasising the popular. On the other hand, Twitter determines what data are available and how data can be accessed through the API (application programming interface). How Twitter provides access to the data in turn affects the analyses the researcher does. The central problem of the thesis is that researchers do not know what relevant data are not collected. Data collection based on keywords, hashtags or users creates data sets that contain fragments of conversations. To solve the problem, a new method was developed. By combining the hashtag and user-based methods, replies to collected tweets were stored, regardless if they contained a tracked hashtag or not.

The four studies this thesis builds on show a complexity of collecting and analysing Twitter data. A key finding is that conversations beyond the hashtag can be quite extensive. As a consequence of this, communication networks based on hashtagged replies were found to be potentially very different from networks based on replies from a more complete data set, where non-hashtagged replies are also included. A network based on hashtagged communication is thus misleading compared to a complete communication network.

Apart from that it is not entirely trivial to identify the parameters to define what should be studied; tests of the API showed that complete data sets cannot be obtained. Therefore, it is important to reflect on both the data collected and the data excluded, not only as a result of the sampling criteria but also what is not given access to. It is also important to be clear about the affordances for interaction that exist when the study is made, both in the user interface but also what API allows and permits.

This research contributes with knowledge about how Twitter is used in the context being studied, but the main contribution is methodological. With the method developed, collection of more complete data sets is enabled, as is analysis of the conversations that take place on the platform. This results in more accurate measurements of the activity. Based on the results of this thesis, there are reasons to suspect that previous studies could differ in terms of results such as communication network size and shape, as well as the type of users that emerges as prominent in the material, compared to if replies that do not contain the studied hashtag had been collected.