Chapter 12
Twitter Data Acquisition and Analysis: Methodology and Best Practice

Stephen Dann
Australian National University, Australia

ABSTRACT

Social media data collection is often treated as tacit knowledge with the collation of tweets reduced to a single sentence without explanation as to means, mechanisms or relative merit of the approach. This chapter describes methods and techniques for the capture of Twitter timeline data, inclusive of first person and third party methods for data capture from personal accounts, public accounts, and keyword searches. The chapter takes a practical approach to acquiring Twitter data with a focus on individual timelines, and small to medium scale search sets. The emphasis is on being able to obtain, examine, and convert Twitter data into knowledge quickly, and with limited requirement for technical skills. This type of data collection assumes no prior programming knowledge. The chapter explains how to retrieve Twitter data from three sources: personally controlled timelines, third party timelines and ongoing search results. Finally, the chapter describes preliminary analysis that can be performed to ascertain content creation patterns, without recourse to analysis of individual tweets.

INTRODUCTION

Collecting Twitter timeline data is a balancing act between the commercial needs of Twitter, the research question driving the data collection, and the assumption of Twitter as a public source of secondary data. The advantage of Twitter over Facebook as a social media pulse is the default public nature of Twitter, versus the default “walled-garden” private nature of Facebook. Twitter accounts default to posting content to the public timeline, and can be viewed from a public website which creates a source of secondary published data. As individual tweets can be attributed directly by URL and citation, it is possible to view an account timeline as a sequence of 140 character standalone publications to be viewed, captured and analyzed as sequential issues in a publically published volume. Twitter generates an extraordinary

DOI: 10.4018/978-1-4666-8408-9.ch012
amount of data as a platform and can be intimidating to consider that this platform can generate in excess of 400 million new tweets per day (Kumar et al. 2014). That said, for studies focused on individual account behaviors or small groups of accounts, there are data collection mechanisms designed to collect small to medium scale data sets. This chapter examines the steps required to obtain timeline data from public Twitter accounts, including a longitudinal search, through a range of capture techniques. It outlines how analysis using just a timestamp and minimal Twitter data can be applied to determine if differences emerge from when, where and what was used to publish to the Twitter account.

BACKGROUND

The research field of Twitter operates at three levels of abstraction for the purpose of data collection – tweet, timeline and pulse. Tweet level is the individually identified message that can be directly accessed by URL. Tweet analysis can take place at the individual level, or within the context of a series of tweets. Timeline is a sequence of tweets from a single account collated directly from Twitter. Time series style tweet analysis to detect patterns in Twitter use over time, and in response to specific external events rely on the timeline. Finally, pulse level data is where Twitter is used to track user sentiment around a keyword, topic or idea across a range of unconnected accounts. Pulse level data is frequently considered the domain of big data using automated analysis and macro-scale capture of millions of tweets. However, this chapter outlines search based tweet capture on specific #hashtags and key terms to provide topic analysis at a large yet manageable bodies of pulse data. Search based data capture also allows for the tracking of brand mentions within the broader Twitterverse or the observation of Twitter engagement with competitor accounts via @mention tracking.

CAPTURING TWITTER DATA: TIMELINE

This chapter features an extended examination of four tweet capture mechanisms to articulate the approaches for collecting data for both industry and academia. Academically, an established method with an explanation of the detail of the acquisition of tweets can be sourced, referenced, and used as a basis to acknowledge variations on collection method. For practitioners, data collection best practice can be used to inform the internal metrics of the organization, or form the basis for a customized protocol to acquire competitor information, or analyze the company’s performance.

The four methods outlined involve internal Twitter account archives, externally mediated timeline capture through Kwitty, keyword search via Hootsuite, and web capture using NCapture as part of NVivo analysis. Each method is discussed in terms of the variables captured in the data, and the steps needed to perform the capture. All methods should be considered equal in their value to a content classification process. Selection and use of a method should be determined by its relative value to an individual project. It may be that Kwitty’s minimalistic four item data set is more valuable for capturing a personally controlled timeline as it will benchmark against subsequent external timeline captures. Alternatively, NCapture’s depth could suit a project requiring greater nuance and pre-prepared coding than is present in the Twitter Archive data. Table 1 outlines a brief comparison of the four methods.
Related Content

E-Commerce Adoption by Micro Firms: A Qualitative Investigation in the UK Tourism Sector
[www.igi-global.com/article/commerce-adoption-micro-firms/54039?camid=4v1a](www.igi-global.com/article/commerce-adoption-micro-firms/54039?camid=4v1a)

Does Instructor's Use of Self-Authored Cases vs. Other Cases in Teaching Lead to More Effective Learning?: Instructor's Use of Self-Authored Cases vs. Other Cases
[www.igi-global.com/article/does-instructors-use-of-self-authored-cases-vs-other-cases-in-teaching-lead-to-more-effective-learning/216126?camid=4v1a](www.igi-global.com/article/does-instructors-use-of-self-authored-cases-vs-other-cases-in-teaching-lead-to-more-effective-learning/216126?camid=4v1a)

Creative Tourism and Cultural Heritage: A New Perspective
[www.igi-global.com/chapter/creative-tourism-and-cultural-heritage/104876?camid=4v1a](www.igi-global.com/chapter/creative-tourism-and-cultural-heritage/104876?camid=4v1a)

Marketing: Looking for a Place in the 21st Century
[www.igi-global.com/chapter/marketing/217094?camid=4v1a](www.igi-global.com/chapter/marketing/217094?camid=4v1a)