Chapter 30
Techniques for Sampling Online Text-Based Data Sets

Lynne M. Webb
University of Arkansas, USA

Yuanxin Wang
Temple University, USA

ABSTRACT

The chapter reviews traditional sampling techniques and suggests adaptations relevant to big data studies of text downloaded from online media such as email messages, online gaming, blogs, micro-blogs (e.g., Twitter), and social networking websites (e.g., Facebook). The authors review methods of probability, purposeful, and adaptive sampling of online data. They illustrate the use of these sampling techniques via published studies that report analysis of online text.

INTRODUCTION

Studying social media often involves downloading publically-available textual data. Based on studies of email messages, Facebook, blogs, gaming websites, and Twitter, this essay describes sampling techniques for selecting online data for specific research projects. As previously noted (Webb & Wang, 2013; Wiles, Crow, & Pain, 2011), research methodologies for studying online text tend to follow or adapt existing research methodologies, including sampling techniques. The sampling techniques discussed in this chapter follow well-established sampling practices, resulting in representative and/or purposeful samples; however, the established techniques have been modified to apply to sampling online text—where unusually large populations of messages are available for sampling and the population of messages is a state of constant growth. The sampling techniques discussed in this chapter can be used for both qualitative and quantitative research.

Rapidly advancing internet technologies have altered daily life as well as the academic landscape. Researchers across disciplines are interested in examining the large volumes of data generated on internet platforms, such as social networking sites and mobile devices. Compared to data collected and analyzed through traditional means, big data generated around-the-clock on the internet can help researchers

DOI: 10.4018/978-1-4666-9840-6.ch030

Copyright © 2016, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

identify latent patterns of human behavior and perceptions that were previously unknown. The richness of the data brings economic benefits to diverse data-intensive industries such as marketing, insurance, and healthcare. Repeated observations of internet data across time amplify the size of already large data sets; data-gathered across time have long interested academics. Vast-sized data sets, typically called "big data," share at least four shared traits: The data are unstructured, growing at an exponential rate, transformational, and highly complicated.

As more big data sets become available to the researchers through the convenience of internet technologies, ability to analyze the big data sets can weaken. Many factors can contribute to a deficiency in analysis. One major obstacle can be the capability of the analytical systems. Although software developers have introduced multiple analytical tools for scholars to employ with big data (e.g., Hadoop, Storm), the transformational nature of big data requires frequent software updates as well as increases in relevant knowledge. In other words, analyzing big data requires specialized knowledge. Another challenge is selecting an appropriate data-mining process. As Badke (2012, p.47) argued, seeking "specific results for specific queries" without employing the proper mining process can further complicate the project instead of helping manage it. Additionally, data of multi-petabyte which include millions of files from heterogeneous operating systems might be too large to back up through conventional computing methods. In such a case, the choice of the data mining tool becomes critical in determining the feasibility, efficiency, and accuracy of the research project.

Many concerns raised regarding big data collection and analysis duplicate concerns surrounding conventional online data collection:

- **Credibility of Online Resources:** Authors of the online text often post anonymously. Their responses, comments, or articles are susceptible to credibility critiques;
- **Privacy Issues:** Internet researchers do not necessarily have permission of the users who originally generated the text. Users are particularly uncomfortable when data generated from personal information, such as Facebook posts or text messages on mobile devices, are examined without their explicit permission. No comprehensive legal system currently exists that draws a clear distinction between publically available data and personal domains;
- **Security Issues:** When successful online posters, such as bloggers, enjoy the free publicity of the internet, they also can be victimized by co-option of their original work and thus violation of their intellectual property rights. It is difficult for researchers to identify the source of a popular Twitter post that is re-tweeted thousands of times, often without acknowledging the original author. Therefore, data collected from open-access online sources might infringe authors' copyrights.

Despite these concerns, researchers and entrepreneurs collect large data sets from the internet and attempt to make sense of the trends contained therein. Howe et al. (2008) issued a call to action for scientists to assist in coping with the complexities of big data sets. Bollier (2010) observed that "small samples of large data sets can be entirely reliable proxies for big data" (p. 14). Furthermore, boyd (2010) raised serious questions about representative sampling of big data sets. Indeed, such incredibly large and complex data sets cry out for effective sampling techniques to manage the sheer size of the data set, its complexity, and perhaps most importantly, its on-going growth. In this essay, we review multiple sampling techniques that effectively address this exact set of issues.

19 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/techniques-for-sampling-online-text-based-data-sets/150187

Related Content

When Should We Ignore Examples with Missing Values?
Wei-Chao Lin, Shih-Wen Keand Chih-Fong Tsai (2017). *International Journal of Data Warehousing and Mining (pp. 53-63).*
www.irma-international.org/article/when-should-we-ignore-examples-with-missing-values/188490

Sociocognitive Inquiry
Brian R. Gainesand Mildred L. G. Shaw (2012). *Social Network Mining, Analysis, and Research Trends: Techniques and Applications (pp. 35-55).*
www.irma-international.org/chapter/sociocognitive-inquiry/61510

Some Aspects of Reliability Estimation of Loosely Coupled Web Services in Clustered Load Balancing Web Server
Abhijit Boraand Tulshi Bezboruah (2020). *Critical Approaches to Information Retrieval Research (pp. 198-209).*
www.irma-international.org/chapter/some-aspects-of-reliability-estimation-of-loosely-coupled-web-services-in-clustered-load-balancing-web-server/237646

Effectiveness of Normalization Over Processing of Textual Data Using Hybrid Approach Sentiment Analysis
Sukhnandan Kaur Johaland Rajni Mohana (2022). *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines (pp. 918-932).*
www.irma-international.org/chapter/effectiveness-of-normalization-over-processing-of-textual-data-using-hybrid-approach-sentiment-analysis/308527

Summarizing Datacubes: Semantic and Syntactic Approaches
Rosine Cicchetti, Lotfi Lakhal, Sébastien Nedjar, Noël Novelliand Alain Casali (2011). *Integrations of Data Warehousing, Data Mining and Database Technologies: Innovative Approaches (pp. 19-39).*
www.irma-international.org/chapter/summarizing-datacubes-semantic-syntactic-approaches/53070