Using the Web as a Data Source: Challenges for Linked Science

Carsten Keßler

Center for Advanced Research of Spatial Information and Department of Geography
Hunter College, City University of New York
New York, USA
carsten.kessler@hunter.cuny.edu

Abstract. The Web makes access to data of interest for disciplines such as geography, sociology, economics, or linguistics almost instantaneous, removing the barrier of lengthy and costly data collection. Using such data from the Web is problematic in terms of the validity, transparency, and reproducibility of the corresponding research, though, as little is known about the subject population and access to the data is often under control of private corporations. This paper discusses the implications of such research and points out potential solutions in the context of Linked Science.

1 Introduction

Data collection has traditionally been the bottleneck of many research endeavors, and gathering the data required to test a hypothesis may still take months, years, or even decades. Physicists need to design and construct complex instrumentation in order to prove the existence of subatomic particles, and astronomers may even have to plan a mission into space to do their work. Other fields, however, are benefiting from an abundance of data at their fingertips, often only a call to a handful of API functions or web services away. Datasets collected through social media platforms such as Facebook or Twitter and collaborative, voluntary efforts such as Wikipedia or OpenStreetMap enable almost instantaneous research across a number of fields. Likewise, the sensor web [7, 3] offers access to an ever-growing amount of real-time data streams. Researchers in geography, sociology, economics, and linguistics, to name but a few, are already using those resources [8, 10, 1, 2, for example].

This discussion paper raises some of the issues that come with the access to those data sources and their use in scientific research. It discusses the implications for validity of the corresponding studies and their reproducibility, and draws conclusions in the context of Linked Science.

2 Validity Issues

Any research that concludes statements about a larger population from a sample is subject to problems caused by a biased sample. Research involving human

subjects can easily find correlations that do not exist at a larger scale if the population investigated is not balanced with respect to the parameters under consideration. As an example, if the goal of a study is to see if there is a correlation between someone's age and the likeliness that they like red wine,[1] the population under consideration must be balanced in terms of other personal and social attributes, such as gender, race, education, and income. If the sample population in our hypothetic study is not chosen appropriately, consisting largely of elder women and younger men, the results might indicate that there is a larger preference for red wine in elder people than in younger people, when in reality, it might be that women prefer red wine more often than men.

This fictional example shows that careful design of the sample population is crucial to come to valid conclusions.[2] Research that draws heavily from social media data, however, is especially prone to this problem, because (a) too little is known about the participants in a study, and (b) the attributes known about the participants are very hard to verify. Moreover, different social media platforms are often used by user groups with different, but distinct, profiles. After all, social media is being used to *socialize* with peers, which will often have at least some demographic properties in common. It is hence difficult to use social media data for studies that are supposedly saying something about the general population. In reality, many of these studies are most likely really only saying something about the users of a specific social media service; this is along the lines of the old joke that many psychology studies do not really say anything about the general public, but a lot about psychology students, as this is the main group taking part in their human participants tests.

3 Reproducibility Issues

Using social media data in research also entails problems for the reproducibility of any studies. User profiles and the data available are in a constant flux, so that it is virtually impossible to replicate a study with the same set of users. While studies that test general statements about the user population of a service (*"Are Facebook users more inclined to conservative political positions than Twitter users?"*) can be replicated in principle, this is only possible as long as the service is available, makes the required information available through its API, and maintains a large user base; all of these factors are outside of the control of the investigator. Finally, data archiving is problematic because in their terms of service, many social networks prohibit making local copies of their data obtained through APIs. Even if such archiving is permitted, the large volume of the data can make archiving difficult or at least expensive.

[1] Clinite [6] suggests that there is no such relationship.

[2] This is a particular pitfall for studies that aim at finding a certain correlation predicted in the hypothesis, and ultimately lead to constructed correlations. The website `http://www.tylervigen.com/spurious-correlations` has some very obvious, yet entertaining examples of such constructed correlations.

Some recent studies have also raised transparency concerns. They have been conducted by some social networks' in-house research teams who had access to data that is not available to anybody outside of the company at that scale [4, 5]. This renders reproducibility completely impossible and forces outside reviewers and researchers to blindly trust the stated results. It also raises the question whether such results should be accepted for publication in the first place. For the studies cited above, the reviewers have decided that the community should know about this research, despite the fact that the basic principle of reproducibility has been violated. The research community will have to come to a consensus for handling such cases as more and more potentially interesting data is collected by private, commercial services that do not provide outside researchers access to their main asset.

4 Implications and Conclusions for Linked Science

Openness, transparency, and reproducibility are core principles of Linked Science [9]. While the scientific community is developing and testing approaches and technologies for data publishing and archiving, they do not work well for research on data from sources such as social media or sensor networks. This kind of data can be hard to archive and make publicly accessible because of the sheer volume, or because of restrictive terms of service of commercial providers. In order to address the latter point, a legislative effort may be required that legalizes data archiving from publicly accessible APIs for research purposes. The scientific community also needs to decide whether it wants to make reproducibility optional, allowing researchers at commercial enterprises to report on findings that no one else can verify or reproduce.

A stricter enforcement of reproducibility and transparency principles for the acceptance of journal and conference submissions is required to solve this problem. The Linked Science principles of semantically annotating, interconnecting, and publishing scientific resources show that the technologies for these processes are already there. Efforts to develop *executable papers* that automatically perform the data analysis steps of a study show that the added value of providing these resources go beyond theoretical reproducibility—they actually reproduce the data analysis. Enforcing the publishing of these resources will also require legal certainty for researchers who work with data from private corporations. As their role as data providers for research is growing, legislation is needed that allows archiving of data obtained from their public APIs. The scientific community hence needs to be more strict about its core principles, leveraging the opportunities offered by technology-driven frameworks such as Linked Science, while the legal circumstances have to be adjusted to ensure transparency without risking lawsuits.

Bibliography

[1] Ballatore, A., Bertolotto, M., Wilson, D.C.: Geographic knowledge extraction and semantic similarity in openstreetmap. Knowledge and Information Systems 37(1), 61–81 (2013)
[2] Benson, E., Haghighi, A., Barzilay, R.: Event discovery in social media feeds. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. pp. 389–398. Association for Computational Linguistics (2011)
[3] Botts, M., Percivall, G., Reed, C., Davidson, J.: Ogc® sensor web enablement: Overview and high level architecture. In: GeoSensor networks, pp. 175–190. Springer (2008)
[4] Burke, M., Kraut, R.E.: Growing closer on facebook: changes in tie strength through social network site use. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 4187–4196. ACM (2014)
[5] Burks, L., Miller, M., Zadeh, R.: Rapid estimate of ground shaking intensity by combining simple earthquake characteristics with tweets. In: 10th US Nat. Conf. Earthquake Eng., Front. Earthquake Eng., Anchorage, AK, USA, Jul. 21Y25 (2014)
[6] Clinite, J.: The Preferences in Wine of Various Aged Consumers. Bachelor thesis, California Polytechnic State University, San Luis Obispo (2013)
[7] Delin, K.A.: The sensor web: A macro-instrument for coordinated sensing. Sensors 2(7), 270–285 (2002)
[8] Ellison, N.B., Steinfield, C., Lampe, C.: The benefits of facebook friends: social capital and college students use of online social network sites. Journal of Computer-Mediated Communication 12(4), 1143–1168 (2007)
[9] Kauppinen, T., Baglatzi, A., Keßler, C.: Linked science: Interconnecting scientific assets. In: Critchlow, T., Dam", K.K.V. (eds.) Data Intensive Science, pp. 383–400. CRC Press, USA (2013)
[10] Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World wide web. pp. 851–860. ACM (2010)