I’m Here, So What? The Value for Statistics of Swiss Geo-Localized Social Media

Daniel Gatica-Perez, Darshan Santani, and Trung Phan Thanh
Idiap and EPFL
Switzerland
{gatica,dsantani,tphan}@idiap.ch

Geo-localized social media from sites like Twitter, Foursquare, and Instagram are being studied in computer science, geography, and urban studies as a source of data to inform urban and mobility phenomena. Geo-tagged tweets and Foursquare check-ins attach physical or semantic location information to content shared online like images and text. This effectively creates links between the physical world and online activity. These data sources have been enthusiastically adopted for their volume and ubiquity due to the popularity of online social networks in developed countries, and are being used both in academia and industry to build new prototypes and services ranging from place and path recommendations to urban monitoring systems.

The work described above has been typically conducted or demonstrated in large cities in the developed world - London, Paris, New York, San Francisco - which have both urban density and significant user bases. Less is known, however, about the specific characteristics of the geo-localized social data produced in Switzerland (freely available from public APIs), as well as its suitability to conduct local urban studies. In a country with a national population of 8 million (about the same as New York City alone), do geo-localized social media data bring any added value for research, compared to other sources of statistical data?

In the presentation, we will summarize a number of empirical studies using Twitter and Foursquare data collected in Switzerland via public APIs over several years, which have aimed to understand a number of urban-related phenomena. More specifically, we have used social data to characterize temporal activity in public transportation hubs; to understand usage patterns of popular city venues; and to characterize patterns of language use in Swiss public venues. On one hand, we will show some promising results that can be obtained from these data sources compared to official statistics, e.g. the estimation of temporal activity in certain types of venues. On the other hand, we will also discuss the different biases that these data sources contain in population and spatial terms, which affect the kind of questions that can be posed and the interpretation that can be done about the results, with emphasis in the Swiss context. Our work aims at developing methodologies in which geo-localized social data complements and enriches official statistics for research in urban and social computing. We will conclude the presentation by making practical recommendations to potential audiences interested in using this type of resources.

Selected References

D. Santani and D. Gatica-Perez, Speaking Swiss: Languages and Venues in Foursquare, in Proc. ACM Int. Conf. on Multimedia, Barcelona, Oct. 2013.
D. Santani and D. Gatica-Perez, Revisiting the Generality of the Rank-based Human Mobility Model, in Proc. ACM Ubicomp Int. Workshop on Pervasive Urban Applications, Zurich, Sep. 2013.
F. Morstatter et al., Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose, In Proc. AAAI Int. Conf. on Weblogs and Social Media, Barcelona, Jul. 2013.
Z. Tufekci,Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls, in Proc. AAAI Int. Conf. on Weblogs and Social Media, Ann Arbor, Jul. 2014.

Target Session: Research and Education, Presentation Format