Myths and Challenges in Knowledge Extraction and Big Data Analysis on Human-Generated Content from Web and Social Media Sources

Marco Brambilla

Politecnico di Milano, Dip. Elettronica, Informazione e Bioingegneria (DEIB).
Via Ponzio, 34/5, 20133 Milano, Italy
marco.brambilla@polimi.it

Abstract. Whatever people produce on digital media can be a relevant source of knowledge and behavioural analysis. This is the subject of interest of a wide part of the new discipline known as Web Science. However, special care must be exercised when setting up studies on this kind of sources. Indeed, these studies rarely satisfy the established scientific method guidelines, because of the nature and size of the data, as well as because of the bias and scarce generalizability of results. This paper identifies some of the most crucial challenges that need to be addressed when tackling knowledge extraction and data analysis out of observational studies on human-generated content.

Keywords: Social Media, Big Data, Data Analysis

1 Introduction

The Web and the social media are today the environments where people post their content, opinions, activities and resources. Therefore, a considerable amount of user-generated content is generated every day for a wide variety of purposes and related to diverse topics and contexts, with high frequency and speed [8].

Exploring and studying this content offers a great opportunity to understand the ever evolving modern society, in terms of topics of interest, events, and people relations and behaviour. Several approaches based on the so called Big Data paradigm have analyzed social media [7, 23, 1, 19, 2], mobile phone usage data [11, 9, 15, 24] and many other sources, with the purpose describe and/or predict behaviour, density, issues, and topics or locations of interests [16, 13]. The potential impact of the analyses can be even larger when information from multiple, diverse sources is merged together [12, 18, 20, 5, 14]. This also include diversity in the format of the data: greater value can be extracted when both textual and quantitative information is analyzed, and even more so when also multimedia content is considered (especially because of the more and more prominent role that photos, video and audio is acquiring in human-to-human communication) [22, 25, 17, 4, 3].

While the data available on this sources is typically created by individual entities, i.e. people or businesses, and pertain their own personal or social sphere, analyses are able to study individuals and also to capture the integrated, high-level view of a more complex ecosystem or phenomenon, such as a city, a country, a community, or an event or social trend.

Therefore, we can say that the aim is to use appropriate tools for understanding complex societal systems and dynamics. Such a tool was conceptualized back in 1976 by De Rosnay [21] who proposed the term *macroscope* as a tool for capturing complex systems. Time is now mature to put this tool at work, by using big data technology and statistical analysis over the huge mass of human-produced content.

In the remaining part of the paper I will present the ideal target of this kind of studies and some risks that are implied by the context and the data collected.

2 The Myth: Comprehensive, Coherent, and Cohesive Analyses

The Big Data paradigm is pushing toward a scenario where people perceive the broad availability of large-scale storage, massive computational power, and flexible and easy to use statistical and machine learning tools. This reflects in the expectation (both in individuals and businesses) to **become able to obtain always and quickly any kind of insights** they may need out of the existing data.

This implies that people expect that the executed analyses will always produce relevant insights for their needs. In particular, results are expected to be:

- **comprehensive**, i.e., always providing a clear picture of the reality with all the relevant aspects, at all levels of details, from very fine-grained results to resilient and solid integrated results;
- **coherent**, i.e., self-consistent in all their versions and dimensions, no matter what is the data source, the granularity of the analysis and the time/space considered;
- **cohesive**, i.e., being able to convey a united and insightful understanding of reality, which can lead to informed and appropriate decision-making processes.

Is this the case? Or, better, is this an automatic outcome that we shall expect from any kind of analysis? Unfortunately, it's quite the opposite, as discussed in the next section.

3 The Challenges of Knowledge Extraction, Web Science, and Content Analytics

In the majority of the analyses, if work is not conducted properly, risks and biases are not considered, and statistics is not applied correctly, **results will suffer of so many critical aspects that they will be hardly instrumental for any objective**. This section has the purpose to introduce some awareness of the risk of big data and knowledge extraction. I don't aim at completeness, but just to report some of the most common challenges that a data scientist need to face when trying to extract knowledge and aggregated analytics out of user-generated content. These challenges include the well known 4 Vs of Big Data, i.e., Volume, Velocity, Variety, and Veracity, and some more. I report them in details below, together with some references to practical experiences.

3.1 Complexity of knowledge

Independently of the technique used for collecting and elaborating the data, and actually also independently of the data itself, we need to keep into account that reality is complex and varies in time, space and along many other dimensions, including societal and economic variables. Aiming at capturing this complexity in its entirety is a goal too ambitious for a single analysis. On the other side, single aspects can be tackled, such as studying the continuously evolving nature of knowledge [10]. Indeed, knowledge in the world continuously evolves, and existing knowledge bases are largely incomplete especially with respect to low-frequency data, belonging to the so-called long tail. Appropriate means can be applied to exploiting content generated on social network platforms, which are excellent sources for discovering emerging knowledge, as they immediately reflects the evolving information and emerging concepts.

3.2 Complexity of data (aka., variety or heterogeneity)

On a more practical level, the data sources that are investigated are complex themselves, as they feature unstructured, multi-format, heterogeneous data. Tackling this variety is challenging *per se*, as each medium and type of content may need different collection, cleaning and analysis techniques. Furthermore, connecting together the extracted insights is challenging too. Finally, an additional complexity aspect is related to the structure of the data, which can feature record formats with data organized in deep hierarchies and relations.

In general, complexity may lead to harder understandability of the data.

3.3 Cognitive bias (i.e., of the observer)

As soon as the data is collected and analysed, an implicit level of bias is introduced. This is the bias of the observer, who is always influencing the way the data is perceived, studied and processed, based on his preliminary knowledge, context, or visibility over the reality. This is also known as *the street lamp effect*: if you look at a completely dark town, with only one lamp on, as an observer you tend to bias your understanding based only on what you see in the lit area.

3.4 Sampling bias (i.e., of the content)

The collected data is usually also biased with respect to the question that the analysis aims to respond. Established scientific method defines the way in which

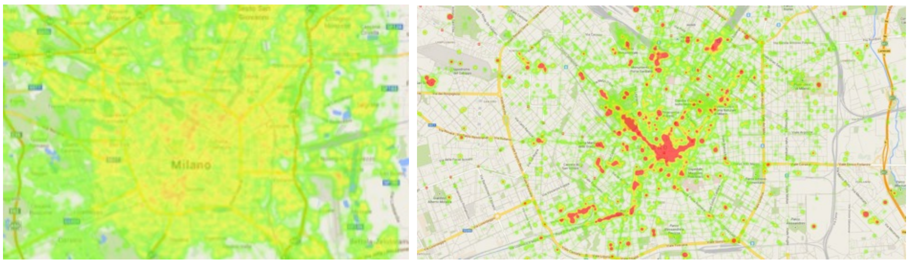

Fig. 1. Two heatmaps representing the density of geolocated photos taken in Milano and shared on two different web sharing platforms, namely Instagram and Flickr.

controlled experiments should be run and conclusions should be drawn. One crucial aspect that should be taken care of is to always apply an appropriate sampling method, granting that the sample is not biased with respect to the population the experiment want to assess. However, this is not so trivial to obtain. For instance, 1 shows two heatmaps representing the density of geolocated photos taken in Milano and shared on two different web sharing platforms, namely Instagram and Flickr: they both feature heavy bias with respect to the total population of the city, due to the mean used for capturing and sharing the photos (mobile app and web site respectively), the audience of the different platforms (casual smartphone users and photographers respectively): the two representations of the city are dramatically different, but there is no "correct" or "wrong" one: it depends on the kind of questions one wants to answer and on the kind of population one is interested in.

Furthermore, in data analysis and extraction from web sources, in many cases controlled experiments may not be implemented. Data scientists frequently need to resort to observational studies, which try to draw conclusions out of observations of the uncontrolled reality of facts and happenings.

3.5 Data Availability

A preliminary problem that many analyses may face is the difficulty in finding and acquiring the data. In many cases, integrated data analysis requires availability of information that is now owned by the analyser himself. Therefore, the analyser need to rely on third parties to release and actually deliver the data. Such third parties may opt for releasing only part of the data, or only on a limited amount per time basis (throttling), or they can simply deliver limited or aggregated views on the data. Some strategies purposefully apply sampling or data limitations in a non-transparent way, so that the data consumer is not able to extract generalizable knowledge out of the analysis work.

3.6 Data Quality

Even when it is actually possible to collect the desired data, a further problem is about the quality of such data. An entire discipline is focused on capturing and solving data quality problems, which can span cases like dummy values, missing values, cryptic data, contradicting data, violation of business rules, data duplication, non-unique identifiers, and many more.

3.7 Data Granularity

Granularity of data refers to the fact that the collected information may refer to different units of analysis on the time scale, space scale, and also along other dimensions. This is a critical issue, especially in the case of analysis on integrated data sources, where each source is based on its own granularity. In this case, techniques based on discretization and uniformation must be applied.

For instance, along the geographic positioning dimension different granularity may mean that some data sources expose data about specific venues/locations, while others expose data about a geographical area defined based on administrative or political borders (e.g., cities, states, provinces, or countries), or on fixed geometrical shapes (squares or Voronoi tassellation); in this case, the solution is usually to rely on a common grid structure. Analogously, along the time dimensions some datasets may provide punctual information, while others may describe different time periods, and thus the solution is to align the data on common periods [6].

3.8 Data Volume

One (possibly obvious) concern regarding data is its volume: in web sources, the amount of data tend to grow enormously. This implies that any analysis needs to cope with the bandwidth, storage, and computation needs of big data. On the other hand, the challenge sometimes is just the opposite: the amount of collectable data is just not sufficient to run sensible analyses.

3.9 Data Coherency

One of the biggest issues with dealing with multiple data sources is that typically the collected data is not coherent: this implies that data cleaning and comparisons procedures must be run, to assess correlation and coherency of the sources.

3.10 Data Velocity (i.e., of the content)

The speed with which the content is produced may also be a concern, as it may become challenging to keep the pace with it in terms of data ingestion, storage and analysis speed. This problem is typically addressed by adopting data streaming solutions when possible.

3.11 Result Velocity (i.e., of the output)

Finally, the speed with which the output is produced is also relevant. Practical experiences show that in most analysis scenarios a strict real-time publishing is not requested. Careful curation of off-line, post-hoc analyses of the phenomenon at large is preferred instead.

4 Conclusion

In this paper I summarized the main expectations and actual challenges in modern data science approaches applied to web and user-generated content. The fact that only observational studies can be applied on web and social network contents is a further complexity factor in these studies. However, if proper data collection, aggregation, cleansing (wrangling) and analysis techniques are applied, these studies can extremely valuable results. Further examples, real cases and discussions can be found in the presentation of this speech, available online.[1]

References

1. Ahmed, K.B., Bouhorma, M., Ahmed, M.B.: Smart citizen sensing: A proposed computational system with visual sentiment analysis and big data architecture. International Journal of Computer Applications 152(6) (2016)
2. Arnaboldi, M., Brambilla, M., Cassottana, B., Ciuccarelli, P., Vantini, S.: Urbanscope: A lens to observe language mix in cities. American Behavioral Scientist 61(7), 774–793 (2017)
3. Bakhshi, S., Shamma, D.A., Gilbert, E.: Faces engage us: Photos with faces attract more likes and comments on instagram. In: Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems. pp. 965–974. CHI '14, ACM, New York, NY, USA (2014), http://doi.acm.org/10.1145/2556288.2557403
4. Bakhshi, S., Shamma, D.A., Kennedy, L., Gilbert, E.: Why we filter our photos and how it impacts engagement. In: ICWSM. pp. 12–21 (2015)
5. Balduini, M., Della Valle, E., Azzi, M., Larcher, R., Antonelli, F., Ciuccarelli, P.: Citysensing: Fusing city data for visual storytelling. IEEE MultiMedia 22(3), 44–53 (July 2015)
6. Balduini, M., Della Valle, E., DellAglio, D., Tsytsarau, M., Palpanas, T., Confalonieri, C.: Social listening of city scale events using the streaming linked data framework. In: International Semantic Web Conference. pp. 1–16. Springer (2013)
7. Becker, H., Iter, D., Naaman, M., Gravano, L.: Identifying content for planned events across social media sites. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. pp. 533–542. WSDM '12, ACM, New York, NY, USA (2012), http://doi.acm.org/10.1145/2124295.2124360
8. Becker, H., Naaman, M., Gravano, L.: Learning similarity metrics for event identification in social media. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining. pp. 291–300. WSDM '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1718487.1718524

[1] https://marco-brambilla.com/2017/09/22/myths-and-challenges-in-knowledge-extraction-and-big-data-analysis/

9. Becker, R.A., Caceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, A., Volinsky, C.: A tale of one city: Using cellular network data for urban planning. IEEE Pervasive Computing 10(4), 18–26 (2011)
10. Brambilla, M., Ceri, S., Valle, E.D., Volonterio, R., Salazar, F.X.A.: Extracting emerging knowledge from social media. In: 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017. pp. 795–804 (2017)
11. Cáceres, R., Rowl, J., Small, C., Urbanek, S.: Exploring the use of urban greenspace through cellular network activity (2012)
12. Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., Ratti, C.: Real-time urban monitoring using cell phones: A case study in rome. IEEE Transactions on Intelligent Transportation Systems 12(1), 141–151 (2011)
13. Candia, J., González, M.C., Wang, P., Schoenharl, T., Madey, G., Barabási, A.L.: Uncovering individual and collective human dynamics from mobile phone records. Journal of physics A: mathematical and theoretical 41(22), 224015 (2008)
14. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: User movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1082–1090. KDD '11, ACM, New York, NY, USA (2011)
15. De Nadai, M., Staiano, J., Larcher, R., Sebe, N., Quercia, D., Lepri, B.: The death and life of great italian cities: a mobile phone data perspective. In: Proceedings of the 25th International Conference on World Wide Web. pp. 413–423. International World Wide Web Conferences Steering Committee (2016)
16. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. arXiv preprint arXiv:0806.1256 (2008)
17. Jaakonmäki, R., Müller, O., Brocke, J.v.: The impact of content, context, and creator on user engagement in social media marketing. In: 50th Hawaii International Conference on System Sciences, HICSS 2017, Hilton Waikoloa Village, Hawaii, USA, January 4-7, 2017 (2017), http://aisel.aisnet.org/hicss-50/da/data_text_web_mining/6
18. Krings, G., Calabrese, F., Ratti, C., Blondel, V.D.: Urban gravity: a model for inter-city telecommunication flows. Journal of Statistical Mechanics: Theory and Experiment 2009(07), L07003 (2009)
19. Psyllidis, A., Bozzon, A., Bocconi, S., Bolivar, C.T.: A platform for urban analytics and semantic data integration in city planning. In: International Conference on Computer-Aided Architectural Design Futures. pp. 21–36. Springer (2015)
20. Quercia, D., Lathia, N., Calabrese, F., Di Lorenzo, G., Crowcroft, J.: Recommending social events from mobile phone location data. In: Data Mining (ICDM), 2010 IEEE 10th International Conference on. pp. 971–976. IEEE (2010)
21. de Rosnay, J.: Le macroscope: vers une version globale. Editions du Seuil (1975)
22. Sabate, F., Berbegal-Mirabent, J., Cañabate, A., Lebherz, P.R.: Factors influencing popularity of branded content in facebook fan pages. European Management Journal 32(6), 1001–1011 (2014)
23. Singh, V.K., Gao, M., Jain, R.: Social pixels: genesis and evaluation. In: Proceedings of the 18th ACM international conference on Multimedia. pp. 481–490. ACM (2010)
24. Wesolowski, A., Eagle, N., Noor, A.M., Snow, R.W., Buckee, C.O.: The impact of biases in mobile phone ownership on estimates of human mobility. Journal of the Royal Society Interface 10(81), 20120986 (2013)
25. Yuheng, H., Lydia, M., Subbarao, K.: What we instagram: A first analysis of instagram photo content and user types, pp. 595–598. The AAAI Press (2014)