The Human Factor in Big Data Analysis

Karl Aberer[1]

[1] *EPFL, 1015 Lausanne, Switzerland*
E-mail : *karl.aberer@epfl.ch*

Abstract

Big Data has become in the recent years central to our information society. Big Data analysis benefits from significant advances in machine learning and data mining. We will demonstrate through concrete examples that human intelligence will remain a crucial factor in Big Data analysis and needs to tightly collaborate with machine intelligence to produce useful knowledge and insights.

Keywords: Big Data, Machine Intelligence, Human-Machine Interaction

1. Introduction

Big Data has become in the recent years central to our information society. Data is being considered as the key economic resource of information society, often compared with the importance of oil for the industrial society. The collection and analysis of massive datasets has become a central factor and driver for transformation for many sectors of industry. Different technological advances are driving these developments, such as increased connectivity, the ability to handle massive data in modern cluster computing centers and the rapidly growing power of machine intelligence due to improvements in algorithms and the ability to analyze increasingly large datasets.

The key challenge in Big Data analysis is the creation of useful knowledge from data. For a long term there has been a disconnection between industry and research, on how this challenge is tackled. Whereas research has been focusing on the use of automated methods based on data mining and machine learning, industry has been relying in business intelligence largely on rule-based approaches, where human experts engineer the rules [Chiticariu et al 2013]. With the rapidly increasing power of machine intelligence, exploited in particular by the big Web companies, e.g., in advertising, this traditional view starts to shift and even turn into quite opposite viewpoints. The need for understanding causal relationships has been proclaimed by some as obsolete, and to be replaced by the pure use of correlations (http://www.wired.com/2008/06/pb-theory/). In other words, human understanding and reasoning would be replaced by the use of statistical patterns.

2. Three Examples of How Human and Machine Intelligence work together

We claim that despite the impressive developments in machine intelligence the use of human intelligence will remain a crucial factor. The key question is not whether the one can replace the other, but how they work together in a productive and efficient way. There are good reasons to believe that human input will remain essential, in particular in areas where domain specific or contextual knowledge is required. In general, machines cannot learn about intentions, valuations and interpretations of humans or human experts, unless massive behavioral traces are collected. This is possible for generic and commonplace concerns (e.g., analyzing Web search logs for optimizing advertisement), but not in settings involving deep domain knowledge. The importance of human intelligence for data analysis is also underlined by the rapid adoption of crowd-sourcing platforms, such as CrowdFlower and Amazon Mechanical Turk, enabling the execution of human intelligence tasks at massive scale.

We will illustrate the interplay between human and machine intelligence through three exemplary recent research works that illustrate different models of interaction.

In the first case we show *how human intelligence is needed in order to enable machine intelligence*. We illustrate this for the problem of evaluating credibility of Web contents, e.g. providing information on health and nutrition [Olteanu et al 2013]. Evaluating credibility relies both on domain expertise, for evaluating factual correctness of information, and on social factors, for establishing common beliefs of what is considered as accepted truth. Ground truth based on human input is primordial to enable any form of automated credibility evaluation. Based on human annotated document corpuses we identified document features that are strong indicators of credibility and serve as the basis for automated credibility evaluation based on supervised learning and recommender algorithms.

In the second case we provide an example of how the computational power of *machines and human intelligence can work interactively together* to solve difficult problems. We consider a task that is of primordial importance in the processing of Big Data, the integration of heterogeneous datasets that originate from different sources [Hung et al 2014]. Using automated analysis of large amounts of structural and content features of database schemas and content allows nowadays coming up with reasonably good suggestions for correlating heterogeneous data, but the correctness of those suggestions is still largely insufficient for high quality data integration. Thus, human expertise and judgment remain an unavoidable component in establishing correct semantic relationships for heterogeneous data. We show how to optimize human interventions, in verifying automatically produced semantic relationships, and thus to significantly speed up the data integration process that is performed in collaborative manner between humans and machines.

In the third example we show of how *human interpretation is needed in order to derive useful*

insights from automatically analyzed Big Data sets for the case of Social Media analysis. Using content and network analysis allows us nowadays to discover a wide range of latent structures in social media data, such as topics, communities and events. Linking such findings to specific business questions remains a challenging task that can only solved by the human experts. We demonstrate for the domain of nutrition, what are possible findings that can be derived from Social Media and the Web through automated analysis with subsequent human interpretation.

3. Conclusion

There is a huge potential in optimizing the interaction of human and machine intelligence. Automated processing of large amounts of data is a great tool for discovering hidden structures and correlations that are inaccessible to humans. However, both for setting the right targets for analysis and for selecting and interpreting the results, humans will remain key in many domains, in particular those requiring specialized expertise such as science and business.

References

Chiticariu, L., Li, Y., & Reiss, F. R. (2013). Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems!. *EMNLP* (No. October, pp. 827-832).

Olteanu, A., Peshterliev, S., Liu, X., & Aberer, K. (2013). Web credibility: Features exploration and credibility prediction. *Advances in Information Retrieval* (pp. 557-568). Springer Berlin Heidelberg.

Nguyen, Q. V. H., Nguyen, T. T., Miklós, Z., Aberer, K., Gal, A., & Weidlich, M. (2014). Pay-as-you-go reconciliation in schema matching networks. *Data Engineering (ICDE), 2014 IEEE 30th International Conference on* (pp. 220-231). IEEE.