Human-Guided Data Mining for Data Journalism

Background

One of the aims of exploratory data mining is to find *unknown unknowns* from data, i.e., discovering facts or insights of which we were previously completely oblivious. As data is collected in many domains, this clearly has many possible applications, both within and outside science. A profession that heavily relies on insights from data is data journalism, which is becoming more and more important in the current *data age*. Insights gained from, e.g., the Panama Papers, Paradise Papers, LuxLeaks, and from the data that Snowden collected have undeniably had a massive impact on society.

Being aware of the importance of data and its analysis, data journalism has become a serious part of many major news organisations. The BBC, for example, has established a dedicated Data Unit. Also, conferences are organised where data journalists and scientists can meet (e.g., SciCAR, https://www.scicar.de/).

Although data journalism appears to be a perfect application for exploratory data mining, in practice very little advanced methods are used. One of the main reasons for this is that many existing methods, including traditional pattern mining algorithms, do not take the knowledge and goals of the analyst into account and, as a result, provide many uninteresting results. Recently proposed methods for interactive pattern mining [1-4] aim to improve this though.

Objective

Investigate and develop human-guided data mining methods for data journalism and apply those to a case study in collaboration with the BBC Data Unit.

More specifically, the goal would be to develop a system that allows data journalists to explore patterns in data and learn from feedback which patterns are relevant and interesting.

Project outline

The steps of a thesis project might be as follows:

1. Study the relevant interactive pattern mining literature;
2. Investigate possible case studies and collect data for one or two case studies, e.g., from https://data.police.uk/data/;
3. Decide on which methods to implement, and probably improve/adapt them;
4. Implement system that enables human-guided pattern exploration;
5. Evaluate system based on simulations and case study;
6. Write thesis (and possibly publication).

Requirements

- We are looking for a highly motivated Master student with good programming skills and experience with data science projects.
- Excellent communication skills.
- Basic knowledge of pattern mining, combinatorial search, and machine learning.

Supervisor

Matthijs van Leeuwen

References

[1] van Leeuwen, M. Interactive Data Exploration using Pattern Mining. In: Holzinger, A & Jurisica, I (eds) *Interactive Knowledge Discovery and Data Mining: State-of-the-Art and Future Challenges in Biomedical Informatics*, LNCS 8401, Springer, 2014.

[2] Dzyuba, V, van Leeuwen, M, Nijssen, S & De Raedt, L. Interactive Learning of Pattern Rankings. *International Journal on Artificial Intelligence Tools* vol.23(6), World Scientific, 2014.

[3] Dzyuba, V & van Leeuwen, M. Learning what matters – Sampling interesting patterns. In: *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining* (PAKDD'17), pp 534-546, Springer, 2017.

[4] Boley, M, Mampaey, M, Kang, B, Tokmakov, P & Wrobel, S. One click mining: interactive local pattern discovery through implicit preference and performance learning. In: *Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics* (IDEA'13), pp27-35, 2013.