Do you find this interesting?

**Principled evaluation of pattern mining algorithms**

*Promotors:* Luc De Raedt, Matthijs van Leeuwen.
*Daily advisors:* Matthijs van Leeuwen, Vladimir Dzyuba.

The goal of this thesis is to develop principled methods for the evaluation of data mining algorithms: are the discovered patterns interesting to the user? Such principles are required for the evaluation of a recent trend towards interactive data mining, where users can influence what is found.

**Problem setting**

Pattern mining is a broad and well-research field within data mining. It is concerned with discovering structure in data, in the form of regularities, where it is important that these are explained by comprehensible descriptions. The best-known pattern mining task is probably frequent itemset mining, which aims at finding sets of “items” that frequently co-occur together, e.g. “bread and butter are often bought together”. Another example is subgroup discovery, which aims at finding patterns that are interesting with respect to a certain target property.

However, the adoption of pattern mining as a primary data exploration tool by domain experts is currently limited, mostly because algorithms mine large numbers of patterns that are often not interesting to the user. For example, top patterns often represent “common knowledge” (“bread and butter are often bought together”). This is due to the fact that algorithms do not take background knowledge and users interests into account. A few algorithms have been recently developed to remedy these issues. However, evaluating the algorithms from the subjective perspective is at least as challenging as designing them. So far it is limited to either emulating plausible real-world scenarios [1, 2], or providing anecdotal evidence via case studies [3].

**Project**

In this project you are invited to develop well-principled methods of evaluating the output of pattern mining algorithms. These methods should provide reproducible and statistically sound results. A very basic workflow might be as follows: a user is presented with the output of several pattern mining algorithms and is asked to evaluate them (“I find Output 2 interesting” or “Output 1 is more interesting than Output 2”). This should allow proving (or disproving) the hypothesis that certain algorithms are better at consistently discovering subjectively interesting patterns than the others.

The methods may leverage insights from statistics, information theory, human-computer interaction (especially information retrieval), and other related fields. Low user effort is an important requirement, therefore tools for effective exploration and comparison of pattern sets (e.g. visualisation of patterns) can also be studied within this project. Existing software might be used as an infrastructure for experiments.

**Profile**

Student with interest in data mining; good understanding of algorithms; good implementation skills.

**Literature**

