Interactive Data Exploration using Pattern Mining

Matthijs van Leeuwen

Machine Learning group
KU Leuven, Leuven, Belgium
matthijs.vanleeuwen@cs.kuleuven.be

Abstract. We live in the era of data and need tools to discover valuable information in large amounts of data. The goal of exploratory data mining is to provide as much insight in given data as possible. Within this field, pattern set mining aims at revealing structure in the form of sets of patterns. Although pattern set mining has shown to be an effective solution to the infamous pattern explosion, important challenges remain. One of the key challenges is to develop principled methods that allow user- and task-specific information to be taken into account, by directly involving the user in the discovery process. This way, the resulting patterns will be more relevant and interesting to the user. To achieve this, pattern mining algorithms will need to be combined with techniques from both visualisation and human-computer interaction. Another challenge is to establish techniques that perform well under constrained resources, as existing methods are usually computationally intensive. Consequently, they are only applied to relatively small datasets and on fast computers. The ultimate goal is to make pattern mining practically more useful, by enabling the user to interactively explore the data and identify interesting structure. In this paper we describe the state-of-the-art, discuss open problems, and outline promising future directions.

Keywords: Interactive Data Exploration, Pattern Mining, Data Mining

1 Introduction

We live in the era of data. Last year it was estimated that 297 exabytes of data had been stored, and this amount increases every year. Making sense of this data is one of the fundamental challenges that we are currently facing, with applications in virtually any discipline. Manually sifting through large amounts of data is infeasible, in particular because it is often unknown what one is looking for exactly. Therefore, appropriate tools are required to digest data and reveal the valuable information it contains.

Although data is everywhere, it is not unusual that the domain experts who have access to the data have no idea what information is contained in it. KDD, which stands for *Knowledge Discovery in Data*, aims to extract knowledge from data. In particular, the goal of the field of *exploratory data mining* is to provide a domain expert as much insight in given data as possible. Although inherently vague and ill-defined, it aims to provide a positive answer to the question: *Can you tell me something interesting about my data?*

As such, its high-level aim is similar to that of *visual analytics*, but the approach is rather different. Whereas visual analytics focuses on visualization in combination with human-computer interaction to improve a user's understanding of the data, exploratory data mining focuses on finding models and patterns that explain the data. This results in (typically hard) combinatorial search problems for which efficient algorithms need to be developed. Depending on the problem and the size of the data, exact or heuristic search is used.

Pattern mining Within exploratory data mining, *pattern mining* aims to enable the discovery of patterns from data. A *pattern* is a description of some structure that occurs locally in the data, i.e., it describes part of the data. The best-known instance is probably frequent itemset mining [1], which discovers combinations of 'items' that frequently occur together in the data. For example, a bioinformatician could use frequent itemset mining to discover treatments and symptoms that often co-occur in a dataset containing patient information.

A pattern-based approach to data mining has clear advantages, in particular in an exploratory setting. One advantage is that patterns are interpretable representations and can thus provide explanations. This is a very desirable property, and is in stark contrast to 'black-box' approaches with which it is often unclear why certain outcomes are obtained. A second large advantage is that patterns can be used for many well-known data mining tasks.

Unfortunately, obtaining interesting results with traditional pattern mining methods can be a tough and time-consuming job. The two main problems are that: 1) humongous amounts of patterns are found, of which many are redundant, and 2) background knowledge of the domain expert is not taken into account. To remedy these issues, careful tuning of the algorithm parameters and manual filtering of the results is necessary. This requires considerable effort and expertise from the data analyst. That is, the data analyst needs be both a domain expert and a data mining expert, which makes the job extremely challenging.

Pattern set mining As a solution to the redundancy problem in pattern mining, a recent trend is to mine pattern *sets* instead of individual patterns. The difference is that apart from constraints on individual patterns, additional constraints and/or an optimisation criterion are imposed on the complete set of patterns. Although *pattern set mining* [2] is a promising and expanding line of research, it is not yet widely adopted in practice because, like pattern mining, directly applying it to real-world applications is often not trivial.

One of the main issues is that the second problem of pattern mining has not yet been addressed: background knowledge of the domain expert is not taken into account. Because of this, algorithms still need to be tuned by running the algorithm, waiting for the final results, changing the parameters, re-running, waiting for the new results, etc. Most existing methods can only deal with interestingness measures that are completely *objective*, i.e., interestingness of a pattern or pattern set is computed from the data only.

Related approaches To tackle the problems of tuning and uninteresting results, Guns et al. [3] advocate an approach based on *declarative modelling*. The analyst can specify the desired results by means of constraints and, optionally, an optimisation criterion. The idea is that constraints are intuitive, and can be iteratively added to the declarative model. A downside is that modelling background knowledge and the task at hand can be tough and still requires substantial skills from the analyst. Furthermore, the constraints need to constructed manually, while interactive approaches could learn these automatically.

Only very few existing exploratory data mining methods use *visualisation* and/or *human-computer interaction* (HCI). To spark attention to the potential synergy of combining these fields with data mining, a recent workshop [4] brought together researchers from all these fields. When visualisation is used in data mining, this is often done after the search [5, 6]. MIME [7] is an interactive tool that allows a user to explore itemsets, but only using traditional interestingness measures, which makes it still hard to find something that is subjectively interesting. Although data mining suites like RapidMiner[1] and KNIME[2] have graphical user interfaces that make data analysis relatively accessible, one needs to construct a workflow and tune parameters.

The importance of taking user knowledge and goals into account was first emphasised by Tuzhilin [8]. More recently De Bie et al. [9, 10] argued that traditional objective quality measures are of limited practical use and proposed a general framework that models background knowledge. This work strongly focuses on *modelling* subjective interestingness, and using the resulting measure for mining is not always straightforward.

Aims and roadmap The purpose of this paper is to discuss the current state-of-the-art in interactive data exploration using pattern mining, and to point out open problems and promising future directions. In the end, the overall goal is to make pattern set mining a more useful tool for exploratory data mining: to enable efficient pattern-based data exploration and identify interesting structure in data, where interestingness is both user- and task-specific.

For this, we argue that it is essential to *actively involve the user in the discovery process*. After all, interestingness is both user- and task-specific. To achieve this, close collaboration between data mining and both human-computer interaction and visualisation will be needed, as Holzinger [11] recently also argued. By integrating efficient pattern mining algorithms into the visual analytics loop [12], and combining these with sophisticated and adaptive subjective interestingness measures, pattern-based data exploration will be able to tell you something interesting about your data.

After providing an introduction to pattern mining and pattern set mining in Section 2, Section 3 describes the current state-of-the-art in interactive pattern mining. After that, Section 4 illustrates the potential of interactive pattern mining with a case study in sports analytics. Section 5 discusses open problems and potential directions for future research, after which we conclude in Section 6.

[1] `www.rapidminer.com`

[2] `www.knime.org`

2 Background and Glossary

This section provides an introduction to pattern mining and pattern set mining, which can be safely skipped by readers already familiar with these areas.

2.1 Pattern mining

Pattern mining aims to reveal structure in data in the form of patterns. A pattern is an element of a specified pattern language $\mathcal{P}$ that describes a subset in a dataset D; a pattern can be regarded as a description of some *local structure*. The commonly used formalisation of pattern mining is called theory mining, where the goal is to find the theory $Th(\mathcal{P}; D; q) = \{p \in \mathcal{P} \mid q(p, D) = \texttt{true}\}$, with q a selection predicate that returns true iff p satisfies the imposed constraints on D.

Many instances of this task exist, with different algorithms for each of them. For example, frequent itemsets [1] are combinations of items that occur more often than a given threshold. In this context, a database D is a bag of transactions over a set of items I, where a transaction t is a subset of I, i.e., $t \subseteq I$. Furthermore, a pattern p is an itemset, $p \subseteq I$, and pattern language $\mathcal{P}$ is the set of all such possible patterns, $\mathcal{P} = 2^I$. An itemset p occurs in a transaction t iff $p \subseteq t$, and its support is defined as the number of transactions in D which it occurs, i.e., $supp(p, D) = |\{t \subseteq D \mid p \subseteq t\}|$. A pattern p is said to be frequent iff its support exceeds the minimum support threshold $minsup$. That is, q returns true iff $supp(p, D) > minsup$, and the theory consists of all itemsets satisfying q. Frequent itemsets can be mined efficiently due to monotonicity of the frequency constraint. Other types of frequent patterns exist for e.g., sequences and graphs.

Subgroup discovery [13, 14] is another example of pattern mining. It is concerned with finding subsets of a dataset for which a target property of interest deviates substantially when compared to the entire dataset. In the context of a bank providing loans, for example, we could find that 16% of all loans with *purpose* = *used car* are not repaid, whereas for the entire population this proportion is only 5%. Subgroup discovery algorithms can cope with a wide range of data types, from simple binary data to numerical attributes and structured data. Subgroup interestingness measures generally compute a combination of the degree of deviation and the size of the subset.

All pattern mining techniques have the disadvantage that the selection predicate q considers only individual patterns. Consequently, vast amounts of similar and hence redundant patterns are found – the infamous *pattern explosion*. Suppose a supermarket that sells n different products. In this case, there are 2^n combinations of products that each form an itemset p. If an itemset p frequently occurs in the data, all $r \subseteq p$ are automatically also frequent. In practice this means that $Th(\mathcal{P}; D; q)$ contains an enormous amount of patterns, of which many are very similar to each other.

An initial attempt to solve this problem was the notion of condensed representations. Closed frequent itemsets [15], for example, are those itemsets p for which no $r \subset p$ exists that describes the same subset of the data. From the set of closed itemsets, the full set of frequent itemsets can be reconstructed and the condensed representation is, hence, lossless. Unfortunately, most condensed representations result in pattern collections that are still too large to be practically useful or interpretable by domain experts.

Also making this observation, Han wrote in 2007 [16]:

> We feel the bottleneck of frequent pattern mining is not on whether we can derive the complete set of frequent patterns under certain constraints efficiently but on whether we can derive a compact but high quality set of patterns that are most useful in applications.

2.2 Pattern set mining

A recent trend that alleviates the pattern explosion is pattern set mining [2], by imposing constraints on the complete result set in addition to those on individual patterns. From a theory mining perspective, this results in the following formalisation: $Th(\mathcal{P}; D; q) = \{S \subseteq \mathcal{P} \mid q(S, D) = \texttt{true}\}$.

Depending on the constraints, in practice this can still result in a gigantic set of results, but now consisting of pattern sets instead of patterns. Mining all pattern sets satisfying the constraints is therefore both undesirable and infeasible, and it is common practice to mine just one. For this purpose, some optimisation criterion is often added. Due to the large search space this can still be quite challenging and heuristic search is commonly employed.

While a pattern describes only local structure, a pattern set is expected to provide a global perspective on the data. Hence, it can be regarded as a (global) model consisting of (local) patterns, and a criterion is needed to perform model selection. Depending on the objective, such criteria are based on e.g. mutual information [17] or the Minimum Description Length (MDL) principle [18]. In all cases, the task can be paraphrased as: *Find the best set of patterns.*

As an example, Krimp [18] uses the MDL principle to induce itemset-based descriptions of binary data. Informally, the MDL principle states that the best model is the one that compresses the data best, and the goal of Krimp is to find a set of patterns that best compresses the data. It was already mentioned that one of the advantages of pattern-based approaches is that patterns can be used for many other data mining tasks. This is particularly true for the compression approach to pattern-based modelling: successful applications include, e.g., classification [19], clustering [20], and difference characterisation [21].

2.3 Glossary

Pattern mining Discovering local structure from data through algorithmic search, where structure is represented by interpretable elements from a pattern language.

Frequent pattern mining Includes frequent itemset mining, but also methods for mining frequent sequences, (sub)graphs, and other pattern types.

Subgroup discovery Subgroup discovery can be seen as an instance of *supervised descriptive rule discovery* [22]. It aims at discovering descriptions of data subsets that deviate with respect to a specified target.

Top-K pattern mining Search for the k best patterns with regard to an interestingness measure. Does not solve the pattern explosion, because of the redundancy in the used pattern languages and correlations in the data.

Pattern set mining Mine *sets of patterns* instead of individual patterns. The large advantage of imposing global constraints and/or having an optimisation criteria is that redundancy can be eliminated.

Descriptive pattern set mining One of the main classes that can be distinguished in pattern set mining, which aims to provide compact and interpretable descriptions of the data.

Supervised pattern set mining A second main class, used when there is a specific target property of interest. Subgroup discovery is an example of a supervised pattern mining task, and pattern set mining variants also exist.

Objective interestingness Almost all interestingness measures for pattern (set) mining up to date are unable to deal with background knowledge or user feedback provided by a domain expert, and are therefore called objective.

Subjective interestingness Interestingness is inherently subjective and should take into account the goals and background knowledge of the current user.

3 State-of-the-Art

For the sake of brevity, in this section we restrict ourselves to recent pattern mining techniques that go *beyond objective interestingness* and enable *user interaction.* We consider both descriptive and supervised techniques. See Kontonasios et al. [9] for a discussion of interestingness measures based on *unexpectedness.*

3.1 Integrating Interaction into Search

Subjective interestingness can be attained in several ways, and one high-level approach is to exploit user feedback to directly influence search.

Bhuiyan et al. [23] proposed a technique that is based on Markov Chain Monte Carlo (MCMC) sampling of frequent patterns. By sampling individual patterns from a specified distribution, the pattern explosion can be avoided while still ensuring a representative sample of the complete set of patterns. While sampling patterns, the user is allowed to provide feedback by *liking* or *disliking* them. This feedback is used to update the sampling distribution, so that new patterns are mined from the updated distribution. For the distribution, a scoring function is assumed in which each individual item has a weight and all items are independent of each other. By updating the weights, the scores of the itemsets and thus the sampling distribution change. Initially all weights are set to 1, so that the initial sampling distribution is the uniform distribution over all patterns.

In similar spirit, Dzyuba & Van Leeuwen [24] recently proposed Interactive Diverse Subgroup Discovery (IDSD), an interactive algorithm that allows a user to provide feedback with respect to provisional results and steer the search away from regions that she finds uninteresting. The intuition behind the approach is that the 'best' subgroups often correspond to common knowledge, which is usually uninteresting to a domain expert.

IDSD builds upon Diverse Subgroup Set Discovery (DSSD) [25]. DSSD was proposed in an attempt to eliminate redundancy by using a *diverse* beam search. For IDSD we augmented it by making the beam selection strategy interactive: on each level of the search, users are allowed to influence the beam by *liking* and *disliking* subgroups, as with the previous method. This affects the interestingness measure, which effectively becomes subjective. IDSD uses a naive scheme to influence the search and, as a result, does not always provide the desired results. However, as we will see in the next section, even a simple method like this can vastly improve the results by exploiting user feedback.

Galbrun and Miettinen [26] introduced SIREN, a system for visual and interactive mining of geospatial redescriptions. Geospatial redescription mining aims to discover pairs of descriptions for the same region, with each description over a different set of features. The system visualises the regions described by the discovered patterns, and allows the user to influence the search in ways similar to those used by IDSD; SIREN is also based on beam search. Although its specialisation to the geospatial setting is both an advantage and a disadvantage, it is another proof-of-concept demonstrating the potential of user interaction.

3.2 Learning user- and task-specific interestingness

Although the methods in the previous subsection use interaction to influence the results, their ability to 'capture' subjective interestingness is limited. This is due both to the type of feedback and the mechanisms used to process this feedback.

Taking these aspects one step further, one can *learn* subjective interestingness from feedback given to patterns. This idea was recently investigated independently by both Boley et al. [27] and Dzyuba et al. [28]. The central idea is to *alternate between mining and learning*: the system mines an initial batch of patterns, a user is given the opportunity to provide feedback, the system learns the user's preferences, a new collection of patterns is mined using these updated preferences, etc. For learning the preferences of the user, standard machine learning techniques can be used, e.g., preference learning. Although the two approaches have a lot in common, there are also some important differences.

The *One Click Mining* system presented by Boley et al. can use any combination of pattern mining algorithms and learns two types of preferences at the same time. One one hand, it uses a multi-armed bandit strategy to learn which pattern mining algorithms produce the results that are most appreciated by the user. This is used to allocate the available computation time to the different algorithms. On the other hand and at the same time, co-active learning is used to learn a utility function over a feature representation of patterns. This utility function is used to compute a ranking over all mined patterns, which is used to determine which patterns are presented, and in what order, to the user. Both learning algorithms completely rely on input provided by means of *implicit user feedback*. Mined patterns are presented in a graphical user interface and the user can freely inspect and store them, or move them to the thrash.

Dzyuba et al. focus on a narrower research question: is it possible to learn a subjective ranking, i.e., a total order, over the space of all possible patterns, from a limited number of small 'queries' that are ranked by a user? For this, they partly build on the work by Rueping [29]. The assumption is that a user has an implicit preference between any pair of patterns, but cannot express this preference relation for all possible pairs. The proposed approach gives the user a small number of patterns (subgroups) and asks her to rank these patterns. RankSVM is then used to learn a preference relation over a feature representation of the patterns, and the resulting utility function can be used to mine subjectively more interesting patterns. An important difference with the approach by Boley et al. is that the learnt utility function is used as optimization criterion in the mining phase, and not only to rank the patterns returned by mining algorithms using objective interestingness measures. Also, query selection strategies inspired by active learning and information retrieval are used to select queries that minimise the effort required from the user.

3.3 Formalising subjective interestingness

All methods discussed so far focus on learning and mining subjectively interesting patterns based on user feedback, but by using a specific learning algorithm they all potentially introduce a strong learning bias. To avoid this, one should first *formalise* subjective interestingness with a principled approach, and then develop the machinery required for using this formalisation.

De Bie [10] has developed a formal framework for exploratory data mining that formalises subjective interestingness using information theoretical principles. The general strategy is to consider prior beliefs, e.g., background information, as constraints on a probabilistic model representing the uncertainty of the data. To avoid introducing any bias, the Maximum Entropy distribution given the prior beliefs is used as model for the data. Given such a 'MaxEnt model', any pattern can be scored against it: one can compute how informative a pattern is given the current model. To avoid overly specific patterns from getting very high scores, the scores are normalised by the complexities of the pattern descriptions.

This framework lends itself well to *iterative data mining*: starting from a MaxEnt model based on prior beliefs, one can look for the subjectively most interesting pattern, which can then be added to the model, after which one can start looking for the next pattern, etc. Because the model is updated after the discovery of each high-scoring pattern, redundancy is avoided. A disadvantage is that the exact implementation of the 'MaxEnt approach' heavily relies on the specific data and pattern types at hand, but instances have been proposed for a variety of data types, e.g. for binary data [30] and multi-relational-data [31].

3.4 Advantages and Disadvantages

Some advantages of pattern-based approaches to exploratory data mining have already been discussed, i.e., patterns are not only interpretable, they can also be

Table 1. Subgroups discovered from the NBA dataset, (a) without and (b) with interaction. Given for each subgroup are its description, its size (number of tuples for which the description holds), and its (objective) interestingness. Taken from [24].

Description	Size	Interestingness
$opp\_def\_reb = F \wedge opponent \neq ATL \wedge thabeet = F$	219	0.0692
$opp\_def\_reb = F \wedge opponent \neq ATL$	222	0.0689
$opp\_def\_reb = F \wedge opponent \neq ATL \wedge ajohnson = F$	222	0.0689
$opp\_def\_reb = F \wedge thabeet = F \wedge opponent \neq PHI$	225	0.0685
$opp\_def\_reb = F \wedge opponent \neq PHI$	228	0.0682

(a) Without interaction – DSSD.

Description	Size	Interestingness
$crawford = F \wedge matthews = T$	96	0.0328
$hickson = T$	143	0.0219
$crawford = F \wedge hickson = T$	63	0.0211
$matthews = T \wedge hickson = T$	99	0.0163
$matthews = T \wedge pace < 88.518$	303	0.0221

(b) With interaction – IDSD.

used for many other data mining tasks. Another advantage is that pattern languages are generally very expressive, which makes it possible to discover almost any local structure that is present in the data. This is, however, also one of the major disadvantages: because the languages are so expressive, in practice many patterns describe highly similar or even equivalent parts of the data.

Specific advantages of the methods presented in this section are that they allow the user to interactively find subjectively interesting patterns, at least to some extent. The methods in the first two subsections are limited when it comes to modelling interestingness, while the MaxEnt approach primarily focuses on scoring patterns and cannot be (straightforwardly) used for interactive learning and/or mining. All methods focus primarily on mining individual patterns rather than pattern sets, although the MaxEnt framework partially solves this by making iterative mining possible. Additional limitations and disadvantages of existing methods are discussed in Section 5.

4 Case Study: Sports Analytics

Let us illustrate the potential of interactive pattern mining with an example taken from Dzyuba et al. [24]. The example concerns a case study on basketball games played in the NBA. More specifically, experiments were performed on a categorical dataset containing information about games played by the Portland Trail Blazers in the 2011/12 season. Each tuple corresponds to a game segment and the attributes represent presence of individual players and standard game statistics. Please refer to [24] for further details.

Table 1 presents the results obtained on this data with two different subgroup discovery methods: one with and one without interaction. As target property of interest, offensive rating was used, i.e., the average number of points per shot. This means that subgroups with high interestingness describe game situations with a high average number of points per shot, which obviously makes it more likely for the team to win the game. The results were evaluated by a domain expert, i.e., a basketball journalist.

For the setting without interaction, DSSD [25] was used with its default parameter settings (Table 1(a)). The results suffer from two severe problems. First, the results are clearly redundant, i.e., diversity could not be attained with the default parameter settings. In fact, the subgroups together describe only 231 of 923 game segments (25.3%). Second, none of the discovered subgroups are interesting to the domain expert, as the descriptions contain no surprising and/or actionable information. For example, it is a trivial fact for experts that poor defensive rebounding by an opponent ($opp\_def\_reb = F$) makes scoring easier, while *absence of reserve players* Thabeet and A. Johnson is not informative either – they more often than not are on the bench anyway.

For the interactive setting, the basketball journalist was asked to use IDSD [24] and evaluate its results (Table 1(b)). With limited effort, he was able to find subgroups that he considered more interesting and actionable: Crawford, Matthews, and Hickson were key players and they often played for the team. So although *objective* interestingness of the subgroups was clearly lower, *subjective* interestingness was substantially higher. In addition, the five subgroups together cover 512 game segments (55.5% of the dataset), implying that the interactive results are also more diverse than the non-interactive. A disadvantage of this particular approach is that not all sessions resulted in interesting results, but this is due to the (ad hoc) way in which feedback is elicited and processed.

5 Open Problems and Future Outlook

We now discuss a number of open problems that we believe need to be solved in order to achieve the overall goal of interactive, pattern-based data exploration.

1. Discovery of pattern-based models for specific users and/or tasks We have argued that purely objective interestingness measures that cannot be influenced are inherently problematic, since interestingness depends on the specific user and task at hand. Hence, adaptivity and subjective interestingness are required. For this, we need an iterative approach to pattern-based modelling that learns what models are interesting during the discovery process, as illustrated in Figure 1. By learning user- and/or task-specific interestingness based on intermediate results, the system can gradually refine and improve its results.

This could be achieved through interaction with a domain expert, but another approach would be to automatically learn task-specific utility, e.g. by having some (automated) feedback procedure as to how useful intermediate results are in an online setting. In such situations an important challenge might be to deal with concept drift, i.e., interestingness must be adaptive and change when needed.

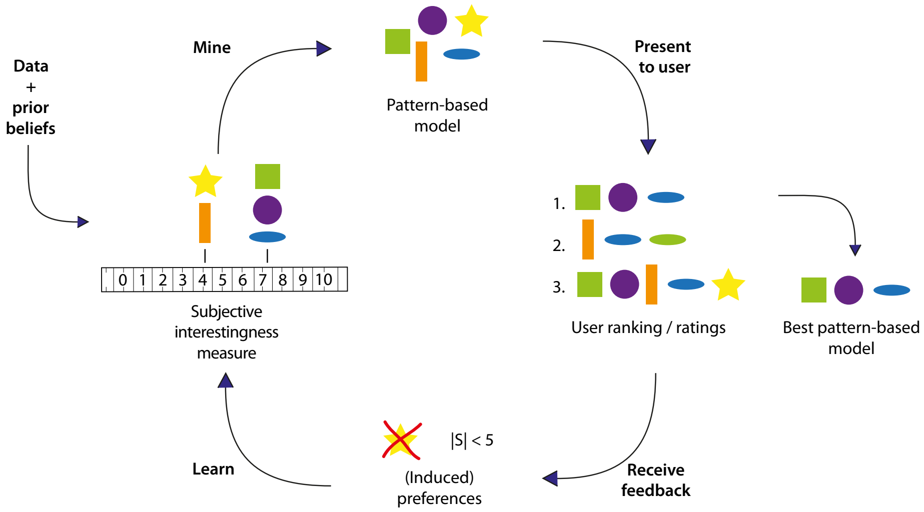

Fig. 1. General approach to learning subjective interestingness for pattern sets.

The methods described in the previous section are a good start in this direction, but they all have their limitations. We need more principled solutions that incorporate both 1) *learning and modelling of interestingness*, and 2) *mining of subjectively interesting pattern-based models*. In particular, most existing interactive pattern mining techniques consider only the subjective interestingness of individual patterns, not that of pattern sets.

2. Resource-constrained pattern set mining through sampling Even on modern desktop computers and dedicated computing servers, existing pattern set mining methods require at least minutes and sometimes hours to compute a result. This makes it hard to apply these methods in a realistic environment, e.g., for purposes of interactive mining, in settings where resources are constrained, or when there is ample of data. Interactive data mining can only become successful if results can be computed and presented to the user virtually instantly.

Pattern (set) sampling according to some subjective, learnt interestingness measure could provide a solution to this problem. Due to extensive redundancy in the solution space, it is sufficient to identify good solutions rather than the optimal solution. These can be presented to, and evaluated by, the user or system, so that feedback can be given and the subjective interestingness can be updated.

3. Principled evaluation of exploratory data mining results Over the past years we have witnessed the publication of a large number of novel algorithms for exploratory data mining. Despite this, there is still a lack of principled methods for the qualitative evaluation of these techniques. Consequently, it is not always clear which methods perform well and under what circumstances.

Although this problem is partly due to the nature of the area, i.e., *exploratory* data mining, the field would greatly benefit from principled evaluation methods.

One approach would be to do (possibly large-scale) user studies, as is also done in information retrieval. It could be argued that the evaluation of pattern sets resembles that of documents retrieved for queries, and therefore measures and techniques inspired by information retrieval could be used. For that reason, collaborations between pattern mining and information retrieval researchers on this topic could be very valuable. A disadvantage is that user studies are complex to conduct, if only because in many cases only one or very few domain experts are available. Another approach might be to construct benchmark datasets for which domain experts know what knowledge they contain. If this can be represented as 'ground truth', this might help to evaluate both existing and novel algorithms. For example, the benchmark datasets made available by the TREC conferences[3] have helped substantially to advance the state-of-the-art in information retrieval.

4. Pattern visualisation for easy inspection and feedback The proposed directions to solving problems 1 and 3 implicitly assume that patterns can be straightforwardly presented to the user, and that the desired feedback can be elicited, but these are non-trivial problems by themselves. To solve these problems, close collaboration with experts from fields like visualisation, visual analytics, and human-computer interaction will be essential.

One problem concerns the *visualisation of patterns together with the data.* Although the descriptions of patterns can be easily presented to a user, interpretation takes time. In particular when a set of patterns is to be evaluated by a user, it would help to visualise the structure in the data that it represents. Even for itemsets and binary data, this can already be quite complex: a single itemset can be visualised as a square in the matrix, but multiple itemsets do not need to be contiguous and may overlap.

A second problem concerns the *interaction between the user and patterns.* Different types of feedback can be used for inducing subjective interestingness, either implicit (inspect, thrash, ignore, etc.) or explicit (ratings, ranking patterns, etc.). But what is the best way to let a user interact with patterns? In the context of pattern mining this question is currently completely unexplored.

6 Conclusions

We argued that it is essential to actively involve the user in the exploratory data mining process in order to discover more interesting results. The state-of-the-art in interactive pattern mining demonstrates that even simple techniques can already vastly improve the results. Still, four important challenges remain.

The first key challenge is to develop principled methods for *learning and modelling user- and task-specific interestingness.* The second challenge is tightly connected to this and is to enable *resource-constrained mining of subjectively interesting pattern-based models.* Once solved, the solutions to these challenges will together form a firm foundation for interactive data mining, but to make this successful the last two challenges will need to be addressed as well.

[3] `http://trec.nist.gov/`

That is, the third challenge concerns the *principled evaluation of exploratory data mining results*, which is important to be able to compare methods. In particular for interactive data mining, solid evaluation methodologies are required, because results are likely to be deemed too subjective otherwise. The fourth and final challenge is to establish *visualisation and interaction designs for pattern-based models*, to enable effective presentation and feedback elicitation.

The ultimate goal is to make pattern mining practically more useful, by enabling the user to interactively explore the data and identify interesting structure through pattern-based models that can be visualised and interacted with.

Acknowledgments. The author is supported by a Postdoctoral Fellowship of the Research Foundation Flanders (FWO). He would like to thank Vladimir Dzyuba for providing useful comments on an early version of this paper.

References

1. Agrawal, R., Imielinksi, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the SIGMOD'93. ACM (1993) 207–216
2. Bringmann, B., Nijssen, S., Tatti, N., Vreeken, J., Zimmermann, A.: Mining sets of patterns: Next generation pattern mining. In: Tutorial at ICDM'11. (2011)
3. Guns, T., Nijssen, S., Raedt, L.D.: Itemset mining: A constraint programming perspective. Artif. Intell. **175**(12-13) (2011) 1951–1983
4. Chau, D.H., Vreeken, J., van Leeuwen, M., Faloutsos, C., eds.: IDEA '13: Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics, New York, NY, USA, ACM (2013)
5. Atzmüller, M., Puppe, F.: Semi-automatic visual subgroup mining using vikamine. Journal of Universal Computer Science **11**(11) (2005) 1752–1765
6. Lucas, J.P., Jorge, A.M., Pereira, F., Pernas, A.M., Machado, A.A.: A tool for interactive subgroup discovery using distribution rules. In: Proceedings of EPIA'07. (2007) 426–436
7. Goethals, B., Moens, S., Vreeken, J.: MIME: a framework for interactive visual pattern mining. In: Proceedings of KDD'11. (2011) 757–760
8. Tuzhilin, A.: On subjective measures of interestingness in knowledge discovery. In: Proceedings of KDD'95. (1995) 275–281
9. Kontonasios, K.N., Spyropoulou, E., De Bie, T.: Knowledge discovery interestingness measures based on unexpectedness. Wiley Int. Rev. Data Min. and Knowl. Disc. **2**(5) (September 2012) 386–399
10. De Bie, T.: An information theoretic framework for data mining. In: Proceedings of KDD'11. (2011) 564–572
11. Holzinger, A.: Human-computer interaction and knowledge discovery (hci-kdd): What is the benefit of bringing those two fields to work together? In: Proceedings of CD-ARES'13. (2013) 319–328
12. Keim, D., Andrienko, G., Fekete, J.D., Görg, C., Kohlhammer, J., Melançon, G.: Visual analytics: Definition, process, and challenges. In Kerren, A., Stasko, J.T., Fekete, J.D., North, C., eds.: Information Visualization. Springer-Verlag, Berlin, Heidelberg (2008) 154–175
13. Klösgen, W.: Explora: A Multipattern and Multistrategy Discovery Assistant. In: Advances in Knowledge Discovery and Data Mining. (1996) 249–271

14. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In: Proceedings of PKDD'97, Springer, Heidelberg (1997) 78–87
15. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Discovering frequent closed itemsets for association rules. In: Proceedings of the ICDT'99. (1999) 398–416
16. Han, J., Cheng, H., Xin, D., Yan, X.: Frequent pattern mining: Current status and future directions. Data Mining and Knowledge Discovery **15**(1) (2007) 55–86
17. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(8) (2005) 1226–1238
18. Vreeken, J., van Leeuwen, M., Siebes, A.: Krimp: mining itemsets that compress. Data Mining and Knowledge Discovery **23**(1) (2011) 169–214
19. van Leeuwen, M., Vreeken, J., Siebes, A.: Compression picks the item sets that matter. In: Proceedings of the ECML PKDD'06. (2006) 585–592
20. van Leeuwen, M., Vreeken, J., Siebes, A.: Identifying the components. Data Min. Knowl. Discov. **19**(2) (2009) 173–292
21. Vreeken, J., van Leeuwen, M., Siebes, A.: Characterising the difference. In: Proceedings of the KDD'07. (2007) 765–774
22. Kralj Novak, P., Lavrač, N., Webb, G.: Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. Journal of Machine Learning Research **10** (2009) 377–403
23. Bhuiyan, M., Mukhopadhyay, S., Hasan, M.A.: Interactive pattern mining on hidden data: A sampling-based solution. In: Proceedings of CIKM '12, New York, NY, USA, ACM (2012) 95–104
24. Dzyuba, V., van Leeuwen, M.: Interactive discovery of interesting subgroup sets. In: Proceedings of IDA 2013. (2013) 150–161
25. van Leeuwen, M., Knobbe, A.: Diverse subgroup set discovery. Data Mining and Knowledge Discovery **25** (2012) 208–242
26. Galbrun, E., Miettinen, P.: A Case of Visual and Interactive Data Analysis: Geospatial Redescription Mining. In: Instant Interactive Data Mining Workshop at ECML-PKDD'12. (2012)
27. Boley, M., Mampaey, M., Kang, B., Tokmakov, P., Wrobel, S.: One Click Mining — Interactive Local Pattern Discovery through Implicit Preference and Performance Learning. In: Interactive Data Exploration and Analytics (IDEA) workshop at KDD 2013. (2013) 28–36
28. Dzyuba, V., van Leeuwen, M., Nijssen, S., Raedt, L.D.: Active preference learning for ranking patterns. In: Proceedings of ICTAI'13. (2013) 532–539
29. Rüping, S.: Ranking interesting subgroups. In: Proceedings of ICML'09. (2009) 913–920
30. Bie, T.D.: Maximum entropy models and subjective interestingness: an application to tiles in binary databases. Data Min. Knowl. Discov. **23**(3) (2011) 407–446
31. Spyropoulou, E., Bie, T.D., Boley, M.: Interesting pattern mining in multi-relational data. Data Min. Knowl. Discov. **28**(3) (2014) 808–849