Useful Patterns (UP'10)
ACM SIGKDD Workshop Report

Jilles Vreeken
Department of Mathematics and Computer Science
University of Antwerp
Belgium
jilles.vreeken@ua.ac.be

Nikolaj Tatti
Department of Mathematics and Computer Science
University of Antwerp
Belgium
nikolaj.tatti@ua.ac.be

Bart Goethals
Department of Mathematics and Computer Science
University of Antwerp
Belgium
bart.goethals@ua.ac.be

ABSTRACT

We provide a summary of the workshop on Useful Patterns (UP'10) held in conjunction with the ACM SIGKDD 2010, on July 25th in Washington, DC, USA. We report in detail on the motivation, goals, and the research issues addressed in the talks at this full-day workshop. More information can be found at: `http://www.usefulpatterns.org`

1. MOTIVATION

Pattern mining is an important aspect of data mining, concerned with finding local structure in data. Traditionally, the focus of research in pattern mining has been on completeness and efficiency. That is, trying to find all potentially interesting patterns as fast as possible. This focus, important as it is, has led our attention away from the most important aspect of the exercise: leading to useful results. To emphasize this, let us consider the following example.

Pattern mining in action, an example

Say a domain expert wants to extract novel knowledge from some data at hand. Or, more specifically, the expert wants to know what patterns are present in the data.

Typically, such data is complex, high-volume, and high-dimensional, and includes a mix of variables that are binary, categorical, hierarchical, or real-valued. Before the expert can apply, say a frequent itemset mining algorithm, the data has to be transformed into a binary matrix. For the numerical attributes, for instance, this involves discretizing the attributes into bins; a non-trivial step, in which potentially important information is easily lost.

Once this conversion is complete, the expert is ready to apply the pattern mining algorithm of choice. Before the mining can commence, however, she first has to define the constraints the patterns need to fulfill, including the main parameter in frequent set mining: the minimal support threshold. Not knowing what the correct value is, she at first sets the threshold at high level. This results in a boring result—the returned patterns mostly represent single items and some trivial associations she already knew. Disappointed by these results, she lowers the threshold somewhat, and starts the algorithm again. Now, in virtually no time at all, a gargantuan number of patterns is returned, together much larger than the original database. To make matters worse, these patterns are typically presented as a text-file.

Nevertheless, let us assume our expert patiently considers the result. Not knowing how to sift through these patterns, she sorts them on frequency, and starts from the top. At first, she sees the same singleton patterns, and as she manually browses the file further she starts to see spurious patterns that can be explained either by singletons or by the trivial associations discovered in the first run. Ideally she would have been given just the most informative patterns. This result should be manageable, and allow the expert to zoom in on particular patterns of interest if needs be.

After arduously considering the many, many discovered patterns, our expert actually finds an interesting pattern, or at least something that is surprising to what is already known. However, the only information typically readily available on this pattern, is the pattern itself, and how often it occurs. To the expert, this is not good enough, as she wants to know *where* the pattern occurs in the data, and whether there is anything interesting happening in those parts of the data, perhaps explaining the pattern, or making it even more interesting. In other words, we need to go back to the data. However, as of yet, there has not been much attention to how this should be done, nor are there tools available to assist in this matter. Further, since the data was transformed, exploring it with regard to a patterns is not trivial.

Making pattern mining useful

All things considered, even when convinced of the potential, in the above case the expert would not be very impressed by the usefulness of pattern mining. Unlike in other fields of data mining, such as clustering, in pattern mining presentation and visualization has not been a priority. However, even when we forget about presentation to a user, patterns are not yet as useful as they could be. While they provide highly detailed descriptions of phenomena in data, it remains difficult to make good use of them in, say, e.g., classification or clustering. While this is mostly due to the huge number of discovered patterns, making the result unwieldy at best, it does pose interesting research questions like 'how to select patterns such that they are useful?'. Techniques that summarize the result exist, but focus primarily on being able to reconstruct the full set, instead of targeting the usability of the summarized set. As such, research into techniques that mine small sets of high-quality patterns is required, where high-quality is directly related their intended use.

In short, it is exactly this kind of research, and these experiences and practices that we discussed at UP.

2. GOAL AND SCOPE

To put it simply, pattern mining is not just taking an off-the-shelf frequent pattern algorithm and applying it to your data. Instead, the actual mining algorithm is just a small part of the discovery process; other tasks are preprocessing data, deciding what type of patterns should be considered interesting to the user, examining results, and analyzing the discovered patterns in the data, and investigating how we can put patterns to good use.

The goal of the Useful Patterns workshop (UP) was to address these problems, and as such, in short, making the results of pattern mining useful. We divided the scope of UP into the following four areas.

Pattern reduction. Our first area of interest is the reduction of the number of returned patterns to useful amounts, whilst retaining the most important information. Techniques to attain this goal can be divided into two main categories.

The first approach is to include background information into discovery process. For example, if we know that both items a and b are frequent, then a high frequency of the combined pattern ab is not very surprising or informative to the end user. Hence, we need to take into account what we already know.

The second category consists of techniques for removing redundant patterns. If we can explain the behavior of a pattern A by pattern B, then we should not report B. Solutions to this end involve analyzing relationships between patterns, as well as investigating how to score patterns as a group, rather than individually.

Presentation. Pattern mining results are typically given in a text file, at best sorted, for example, by their frequency. Obviously, exploring billions, or even just hundreds, of patterns manually is too laborious. By visualizing the patterns and allowing user to explore them we can greatly reduce the task of interpreting the results: the user will get the big picture instantaneously, while at the same time can explore the discovered patterns in greater depth.

Presentation is not limited to visualization, as an important aspect of this topic is to be able to regard a pattern in the original data; allowing the expert to analyze whether the pattern is true or spurious, novel or well-known, and most importantly, whether it is worth further investment of further effort.

Together, research on the presentation of patterns will make pattern mining a better tool for exploring data, and as such, allow it to be used in practice by experts.

Using patterns. Our third main topic was to discuss how groups of patterns can be used as surrogates for data. The main idea here is to challenge the belief that patterns are the end goal, but instead use them as an intermediate result. By replacing data with patterns, we have transformed information in the data into a different form. This representation may be more suitable for other algorithms such as clustering, classification, etc. As highly efficient pattern mining algorithms have been developed, these surrogates may be used to analyze very large data collections by techniques that could otherwise not consider the full database.

Use cases in pattern mining. Discovering patterns can be a difficult process if the original data doesn't fit perfectly into the pattern mining setup. The goal of this topic was to discuss how to mine patterns from real-world data, ranging from preprocessing to defining new pattern types. Clearly, this is a very important area of research, if pattern mining is to be used, and *useful*, in practice by experts.

3. OVERVIEW

UP 2010 was the first workshop on making pattern mining *useful*. The program of UP consisted of two invited keynote talks of 45 minutes each, and eight 20-minute regular research presentations. As organizers, we strived to make the program diverse and engaging, attending to the main topics identified above. All presentations are available for download from our website, `http://www.usefulpatterns.org`.

4. INVITED TALKS

We are proud to have had two excellent invited talks at the workshop.

The first invited presentation of the workshop was a keynote titled 'Mining Useful Patterns: My Evolutionary View', given by Jiawei Han from the University of Illinois at Urbana-Champaign. In this talk prof. Han detailed how the ideas on how we should mine patterns that are useful have evolved since the conception of pattern mining; starting our more than a decade ago by simply finding frequent sets, to the current state of the art where patterns and semantic annotation are combined to provide deep and useful insight. An extensive number of examples of what types of patterns are considered to be useful in certain practice were given. With these examples and practices in mind, a number of open research problems and less-well explored areas within pattern mining were pointed out, all of which will offer great opportunities to make pattern mining more useful.

Geoff Webb from Monash University gave an excellent keynote presentation on the topic of 'Association Discovery'. In the talk Geoff introduced association discovery, and expounded on how it differs from what statisticians do in traditional correlation analysis. A major topic of the talk was top-most interesting patterns, and how to find these—discussing the strengths and limitations of using statistical testing to do so. The point the talk drove to, was the question what we should look for: associations, association rules or itemsets. The answer, according to Geoff, are itemsets. Further, in order to find the best itemsets, pattern miners should learn from statistics in order to return only the most interesting patterns, and so avoid reporting redundant information.

5. PRESENTED RESEARCH

As our example in the introduction mentions, **data preparation** is an important step before one can apply pattern mining. In their paper titled 'Multi-Resolution Patterns from Binary Data', Prem Adhikari and Jaakko Hollmén explore approaches to sample features with regard to pattern mining, inspired by the multi-resolution availability of DNA amplification data. Besides potential scalability issues, data gathered at very high resolutions is much more likely to contain (high levels of) noise, and such lead to spurious patterns. While much simpler to mine, very low resolution data,

on the other hand, may not include important patterns. In this work, the authors investigate how high-resolution data can best be sampled in order to reduce levels of noise, but yet keep the important patterns present.

One of the main areas of interest for the workshop was that of **pattern reduction**, and at the workshop this area received ample attention with four presentations.

A conceptual framework to mine interesting patterns was proposed by Tijl De Bie, Kleantis-Nikolaos Kontonasios, and Eirini Spyropoulou. In the presentation, Tijl De Bie detailed how when mining for useful patterns, we should take the background beliefs of the expert into account, and how we can do this theoretically and practically. By infusing background beliefs, such as notions of particular structure and/or associations, into a Maximum Entropy model of the data, the significance of patterns can be tested straightforwardly. Experiments using this framework to discover tiles in binary data show the approach to work well in practice.

Similar in vein, with regard to scoring individual patterns, Anne Denton presented work together with Jianfei Wu and Dietmar Dorr titled 'Point-Distribution Algorithm for Mining Vector-Item Patterns'. In the presentation, Anne explained how each transaction is assumed to have both some binary features, as well as a continuous attribute. The key aspect of their algorithm is that it returns patterns for which the distribution of the continuous attribute(s) follows the presence of an itemset.

A major topic in the reduction of the number of patterns, is the condensation of a large set of patterns into a smaller, yet (almost) equally informative one. The presentations by Jin et al. and Fradkin and Moerchen both addressed this particular sub-topic, but did so with two very different approaches. The former presented a paper titled 'Block Interaction: A Generative Summarization Scheme for Frequent Patterns', in which blocks in the data are identified that give rise to particular patterns. Given these blocks, most, if not all, those patterns can be reconstructed with high precision. The experiments show that this mining technique leads to only handfuls of discovered patterns. The latter focused on sequential patterns, and expands the notion of the well-known closure operator to that of approximate-closure. In this paper, titled 'Margin-Closed Frequent Sequential Pattern Mining', Fradkin and Moerchen extend the BIDE miner such that it does not report those sequential patterns that have almost they same frequency than their super-patterns. Experiments show that already for small error-margins very large reductions in the number of patterns are attained.

On the topic of **pattern presentation**, Michael Carmichael and Carson K. Leung presented a visualization technique in their paper titled 'Visualising Useful Patterns'. In their approach, the authors visualize itemsets in a plot with x-axis representing items and y-axis representing the frequency. An itemset is then plotted as a series of connected circles. To avoid clutter, authors use several techniques: they use only closed itemsets, itemsets with same the frequency are collapsed into one set, these groups can be dynamically expanded to show individual items, and finally itemsets having the same prefix and the same frequency are grouped into trees. By doing so, the authors framework allows the user to see the big picture instantly while at the same time providing means for more detailed exploration.

On our key topic of **use cases**, the paper of Arne Koopman, Arno Knobbe, and Marvin Meeng, described how pattern mining can be used to monitor the the structural integrity of important infra-structure, in their case a bridge. This bridge, the 'Hollandse Brug' in the Netherlands, has been outfitted with a large number of sensors, that together provide very high-resolution data of the structural integrity. By using pattern mining, the large stream of data can be characterized, and degradation can be detected, as the patterns for the normal state would not fit. In their submission, titled 'Pattern Selection Problems in Multivariate Time-Series using Equation Discovery', the authors describe equation discovery can be employed to find patterns in these high-resolution multivariate time-series. The work is currently ongoing, and the speaker announced that (part of) the data will likely be made available at a later date.

On the topic of **using patterns**, Kim et al. presented a technique for determining the authors of a given paper, in their paper titled 'Authorship classification: a syntactic tree mining approach'. In their approach, the authors construct a syntactic tree describing the grammar structure in the given text document. The authors continue by discovering patterns, frequent subtrees, from this tree. The most discriminative patterns are selected and fed as features to an SVM classifier. The authors demonstrate empirically that these features outperform other syntactic features, such as, function words, POS tags, and rewrite rules.

6. WORKSHOP ORGANIZATION

Workshop Co-Chairs
Bart Goethals, University of Antwerp
Nikolaj Tatti, University of Antwerp
Jilles Vreeken, University of Antwerp

Program Committee
Michael Berthold, University of Konstanz
Björn Bringmann, K.U. Leuven
Johannes Fürnkranz, T.U. Darmstadt
Vivekanand Gopalkrishnan, Nanyang Tech. University
Ruoming Jin, Kent State University
Eamonn Keogh, University of California – Riverside
Arno Knobbe, Universiteit Leiden
Arne Koopman, Universiteit Leiden
Carson K. Leung, University of Manitoba
Srinivasan Parthasarathy, Ohio State University
Jian Pei, Simon Fraser University
Kai Puolamäki, Aalto University
Geoff Webb, Monash University

All submissions were considered by typically three different reviewers. The workshop co-chairs selected the papers to be presented at the workshop, and included in the workshop proceedings, based on these reports. The proceedings are available online through the ACM Digital Library. More information at: `http://www.usefulpatterns.org`

Acknowledgements

We wish to thank the program committee for their excellent and professional input, making the selection process very light work. We also wish to thank the keynote speakers, Jiawei Han and Geoff Webb, for giving engaging and thought-provoking talks. A special thanks goes to the authors of all submitted papers, as their excellent contributions allowed us to set up an engaging workshop program. Last, but not least, we thank all attendees.