MIME: A Framework for Interactive Visual Pattern Mining

Bart Goethals   Sandy Moens   Jilles Vreeken

ADReM
University of Antwerp
{first.lastname}@ua.ac.be

ABSTRACT

We present a framework for interactive visual pattern mining. Our system enables the user to browse through the data and patterns easily and intuitively, using a toolbox consisting of interestingness measures, mining algorithms and post-processing algorithms to assist in identifying interesting patterns. By mining interactively, we enable the user to combine their subjective interestingness measure and background knowledge with a wide variety of objective measures to easily and quickly mine the most important and interesting patterns. Basically, we enable the user to become an essential part of the mining algorithm. Our demo[1] currently applies to mining interesting itemsets and association rules, and its extension to episodes and decision trees is ongoing.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining

General Terms

Algorithms, Experimentation

1. INTRODUCTION

Data mining is an inherently iterative process; the results of one analysis often lead to new questions, requiring more analysis. In an ideal world, this process is streamlined. That is, data mining is not only iterative, but also interactive: the user can give such feedback immediately, and easily browse the results. In traditional pattern mining, however, algorithms typically produce large amounts of patterns, many of which are not interesting to the user [9], and the results are typically only given in a flat text file, making it hard to analyze the results. By instead providing an iterative and interactive process, the user would be able to explore and refine the discovered patterns on the fly.

In this paper, we present a framework in which we allow the user to interactively mine a database for interesting itemsets and association rules. Our system visualizes all patterns discovered so far, yet, importantly lets the user interactively explore and dynamically modify these in an intuitive manner [2, 10]. By the visualization, users can browse through the mined data more easily and so quickly discover important information. Essentially, in our framework visualization strengthens the interactive mining process, and vice versa.

Basically, our framework, MIME (Making Interactive Mining Easy), draws from user knowledge and interest to improve the collection of patterns discovered by the mining algorithms, by letting the user take control during the mining process, and allowing to adapt the results and so create useful collection of patterns. To assist the user, MIME offers an extensive toolbox of interestingness measures, mining techniques, and visualizations. Using these, the user can easily identify and remove uninteresting or redundant patterns, extend or reduce existing patterns, or apply various post-processing techniques. As such, in our framework the user becomes an essential part of the mining algorithm.

2. DESCRIPTION OF THE SYSTEM

We consider transactional (supermarket) databases where each transaction contains a number of items. A pattern is an itemset (a group of items $\{A, B, C\}$ that occur together) or a rule (two groups of items $\{D, E\} \Rightarrow \{F, G, H\}$ where the presence of the first group implies the presence of the second group with a given confidence). The support s of a pattern is the number of transactions in the database in which the pattern is present. The frequency f is support relative to the size of the dataset. In this setting frequent itemset mining is defined as the process of finding all patterns in a database that have a frequency f_p above a user-specified threshold f.

Most pattern mining techniques produce an amount of output that due to its size is difficult to post-process. One could try to reduce the number of results by making the quality thresholds more strict. Unfortunately this does not guarantee the usefulness of the produced patterns. To this end more is required.

MIME combines knowledge of the expert and computational strength of a computer to increase the probability of finding interesting patterns. This is achieved in a visual framework where the expert can create his own pattern collection. In order to evaluate the created/mined patterns, MIME implements many interestingness measures. This way the user can select a number of measures that are important for his purpose/domain. The measures improve the knowledge of the user. Our tool also contains a number
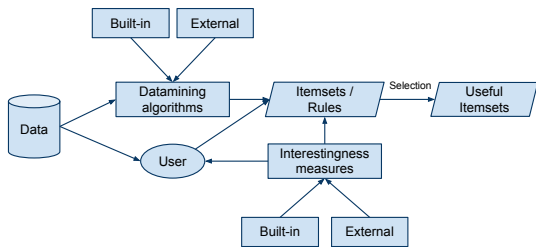
[1] http://adrem.ua.ac.be/mime

Figure 1: MIME Workflow

of standard mining and post-processing algorithms such as Eclat, Apriori, etc. The mining algorithms can be used to construct a starting set of patterns with which a user can play and further construct his collection of useful patterns.

The overall workflow of MIME is shown in Figure 1. Data is read from a database file and can be used by a user or by mining algorithms to create rules and itemsets. From the created patterns a user can make a decision of which patterns are useful, based on the provided interestingness measures. The user may also apply post-processing algorithms to the created collection of patterns.

All information provided by MIME is computed on-the-fly. When selecting new measures from the toolbox, it computes the corresponding ranks of the discovered patterns and shows them as soon as they are available. Extending or reducing patterns also reflects in an immediate recalculation of the applied measures. The nature of any-time algorithms applies to our framework in the sense that partial information is shown when available.

An important feature is *Best Pattern Extension*. This feature allows the user to immediately see the individual impact of remaining items on an existing pattern, i.e. a pattern is extended separately by all remaining items in the dataset and the resulting ranks (based on the active measure) are computed and shown in the source dock, again using rank decreasing order. Other functionality for generation of patterns based on existing patterns, are the generation of all subsets, subsets of given size, closed supersets etc. Mining algorithms can also use discovered patterns to filter the original result. All these features (including mining algorithms and post-processors) provide the user with a toolbox for easy and fast generation of possibly useful patterns.

From the workflow also notice the *External* nodes indicating the use of external/plugin functionality in the tool. In order to provide a widely applicable and easily extensible tool, we have equipped our tool with a plugin system, such that existing mining and post-processing algorithms can be used. The produced patterns are automatically loaded into the MIME framework.

3. MIME

MIME contains a number of different panes a user can interact with. The panes are shown in Figure 2. They are interconnected such that changes in one pane have an effect on one on the other panes. The different panes are:

- *Source Dock*: contains items (the building blocks for rules and itemsets) that are present in the dataset that has been loaded into the framework. All items are automatically sorted in rank decreasing order. The ranks are based on the active measure and possibly

a selected pattern. This view also allows grouping of items in the dataset in an item taxonomy. This grouping can be useful in supermarket databases where for instance coke bottles with different sizes represent different items in the data. A limiter has also been added such that the number of visible items can be reduced, for instance based on minimum support.

- *Work Dock or Workspace*: shows the constructed and mined patterns. To construct patterns, items can be dragged from the source dock or from already existing patterns. The work dock also shows the interestingness of patterns in two possible manners, using numerical values or using a coloring scheme. An example is shown in Figure 2(b). Also, the number of workspaces is not limited to one; different spaces (collections of patterns) can be created and compared, and information can be sent from one space to another.

- *Toolbox*: shows built-in and plugin measures, mining algorithms and post-processing algorithms that can be used to interact with the data from the source and the work dock. Typically, a user starts by selecting a few measures from the measure bar. The user then runs a mining algorithm or creates a few patterns of its own (possibly based on the output of a mining algorithm) and then runs a post- processing algorithm that is applied to the mined patterns.

- *Global Overview*: an overview of all workspaces is shown here. Information such as the number of items and the number of patterns, are shown by default. Built-in and plugin measures can be used here to compare workspaces. It is thus also possible to compare the outcome of multiple algorithms by applying these algorithms on the same dataset in different workspaces.

4. IMPLEMENTATION

The tool has been implemented in Java using the Qt-Jambi[2]-library for visualization and GUI-purposes.

MIME provides a framework for the exploration of patterns, where user can create a pattern, see the best pattern extension, add or remove items and so obtain useful patterns. As calculating some measures may be computationally expensive, caching has been used to minimize the number recalculations. Threading has been used to compute as much as possible in the background, not slowing down the user's ability to further explore the dataset. The combination of these two techniques in the tool, creates an environment where interactivity remains high.

A plugin framework has also been used in order to make our tool easily expandable. Plugins can be defined using the configuration file. For the moment there is one constraint on the plugins that can be used: it must contain command-line parameters for an input and output file, which are specified by position of the parameters.

5. ILLUSTRATIVE SCENARIO

In this section we give a demonstration of the capabilities of our tool using the stemmed version of the *ICDM Abstracts* dataset by Tijl De Bie [6]. The database contains
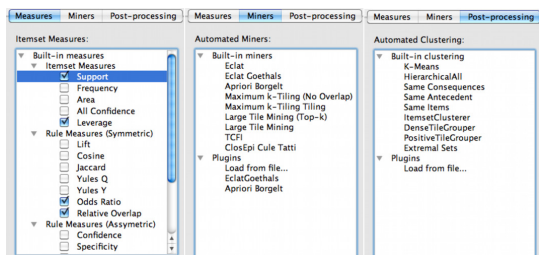
---

[2]http://qt-jambi.org/

(a) Source dock

(b) Work dock

(c) Toolbox

(d) Global overview

Figure 2: MIME Views

Figure 3: *Best Pattern Extension* $\{\} \Rightarrow \{larg\}$

best pattern extension is activated (Figure 3). This way we can easily see which items are potentially interesting for an existing itemset. For the itemset *larg*, 3 items appear to produce a good ranking (above 0.0250). We select those items and create three new itemsets by dropping the items inside the basket and choosing to create distinct rules. We can then evaluate the remaining patterns in the same fashion, adding new items as long as the quality remains good. The quality is subject to the active measures and to the knowledge of the expert. Figure 2(b) shows a pattern collection that is the result of this process.

We can also apply a mining algorithm to the dataset. We start by selecting a few patterns (this is not mandatory), and use the apriori-implementation by Borgelt[3]. Parameters can be specified and the algorithm can be executed. By selecting patterns from the work dock, the outcome of the algorithm is automatically filtered, i.e. the resulting patterns (if any) are supersets of one of the selected patterns. We may now also apply a post-processing algorithm to the collection of patterns, for instance positive tile clustering.

Rules can be mined in similar fashion, dragging items to the left (resp. right) side of an existing itemset creates a body (resp. head) for this itemset. We can now select the left hand side or the right hand side of the rule to get a reordering of the remaining items in the dataset.

Worksheets and comparing miners

Our tool allows for multiple worksheets. Worksheets can be based on a dataset that is already in use, in which case rules can be copied or moved from one sheet to another. It is also possible to use one of the mining algorithms to construct a collection of patterns, select some of the patterns and copy these to a new worksheet such that we do not lose the original mining results.

When using different worksheets, we can also use the global overview and global measures to compare the results of different mining sessions of the same algorithm, or to compare the results of different mining algorithms.

For the first comparison we create 4 different workspaces using the *Abstracts* database, and run the Apriori miner by Borgelt using 4 different frequency levels (40, 30, 20 and 10%). Figure 2(d) shows the global overview of this setting. It is clear that more itemsets provide a greater coverage, however we can now adapt the results of the mining algorithms to improve these values.

For the second comparison we can for instance compare the outcome of 4 different tile-based methods with similar thresholds, this way the user can see which miner gives better results for a certain dataset, using information provided by the global overview as reference.

Worksheet colors and basket colors

MIME supports graphical and textual (numerical) values for showing ranks. The numerical values can be used to obtain the worksheet coloring shown in Figure 2b. As is the case

---

[3]http://www.borgelt.net/apriori.html

4976 items over 859 transactions. Each transaction represents an abstract of a paper that was published between 2001 and 2008. We start by loading the dataset in memory using the worksheet menu.

Grouping items

In this first step we show how to create item hierarchies. We can select items by clicking on them in the source dock while holding the control modifier. We then use the group button to create a group. For instance grouping together the items *database* (support of 139), *dataset* (70) and *datasets* (132) into a new item called *database-group* (287). The group takes the union of the transactions of the individual items. The new item can now be used as all normal items and is now used when running mining algorithms on the database.

Mining itemsets and rules

To demonstrate how to mine itemsets, we use items with average support (*larg*, *learn* and *time*), and select a few extra measures. We set the leverage measure as primary measure by double clicking.

When selecting the first itemset by clicking on the basket,

with the ranking of the items in the source dock, patterns can obtain a coloring scheme depending on the interestingness measure and the other rules in the worksheet. This way the user gets a quick understanding of which measures are similar for patterns without knowing the exact ranks.

Items inside a basket of a rule can also obtain a coloring scheme based on their personal influence on the pattern. The ranking of the pattern is computed when omitting the specified item and the rank is compared to the current ranking. If the current ranking is better, the item gets a positive coloring (green), and if the current ranking is worse, the item gets a negative coloring (red). If the item does not have an influence it gets a neutral (white) coloring.

6. RELATED WORK

A lot of work has been done comparing and evaluating different objective interestingness measures [8, 11]. The most important outcome is that there is no single measure that can be used for all purposes, and even worse, for some purposes, there exists no measure, and only subjective and semantic criteria based on experience and background knowledge can produce actionable results.

In our tool we have incorporated several objective interestingness measures, but it is the combination of user-knowledge and objective measures enabling subjective interestingness criteria to be applied.

Also in the context of Inductive Databases several interactive constraint-based mining frameworks have been studied [7]. Here, the user typically has the ability to specify all kinds of constraints that are being used during the mining process. In these systems, the visualization and adaptation of the results has not been a major concern. Moreover, our framework could be built on top of such an inductive database implementation [3].

Most similar to the system presented here is the framework for mining decision trees proposed by Ankerst et al. [1]. The user and computer work together in this process such that in each step either the user or the computer can make the decision for a new split. The computer also provides extra computational power by showing the best split, lookahead information, purity measures etc. Our approach is similar, but applies to frequent itemsets and association rules, instead of decision trees.

Some methods for visualizing patterns exist [4, 5]. These visualizers present the output of mining algorithms in a compact and graphical format, and allow to further filter the output using queries. They do not provide means to mine the database interactively using subjective criteria and also do not allow to further explore existing patterns.

7. CONCLUSIONS AND FUTURE WORK

MIME is a framework for the interactive mining, exploration, and post-processing of patterns, and allows for easy comparison of different algorithms and pattern collections. It makes the user part of the mining process and so allows for creating patterns and adapting patterns by different mining algorithms and quality measures, as well as personal knowledge and interest. Post-processing algorithms (clustering algorithms for instance) can be applied to the created collection or the user can create hierarchies himself. The tool also contains a plugin system that allows extension of the tool with existing software.

There are still many options for extending MIME that we consider important future work. First, iterative mining algorithms can be incorporated in this system, allowing a speed- up of mining sessions with different quality constraints. Second, caching and pre-computation can be used in this framework by keeping track of user-interactions that are performed most and creating an interaction hierarchy. The system learns about user- interactions and can then decide to pre-compute important calculations. Third, support for mining of decision trees and episodes is in progress. At last, we can also make more use of any-time algorithms, for instance when computing the *Best Pattern Extension*.

8. ACKNOWLEDGMENTS

9. REFERENCES

[1] M. Ankerst, M. Ester, and H.-P. Kriegel. Towards an effective cooperation of the user and the computer for classification. In *Proc. ACM SIGKDD*, pages 179–188, 2000.

[2] E. Bertini and D. Lalanne. Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery. *SIGKDD Explor. Newsl.*, 11:9–18, May 2010.

[3] H. Blockeel, T. Calders, E. Fromont, B. Goethals, A. Prado, and C. Robardet. An inductive database prototype based on virtual mining views. In *Proc. ACM SIGKDD*, pages 1061–1064, 2008.

[4] C. L. Carmichael and C. K.-S. Leung. FpVAT: a visual analytic tool for supporting frequent pattern mining. *SIGKDD Explor. Newsl.*, 11:39–48, May 2010.

[5] A. Datta and K. Techapichetvanich. VisAR : a new technique for visualizing mined association rules. In X. Li, S. Wang, and Z. Dong, editors, *Adv. Data Min. Appl.*, pages 728–728. Springer Berlin / Heidelberg, 2005.

[6] T. De Bie. Maximum entropy models and subjective interestingness: an application to tiles in binary databases. *Data Min. Knowl. Disc.*, In press.

[7] S. Dzeroski, B. Goethals, and P. Panov. *Inductive Databases and Constraint-Based Data Mining*. Springer-Verlag, 2010.

[8] L. Geng and H. J. Hamilton. Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, 38, September 2006.

[9] B. Goethals, N. Tatti, and J. Vreeken. Useful patterns (UP'10) ACM SIGKDD workshop report. *SIGKDD Explor. Newsl.*, 12:56–58, March 2011.

[10] D. A. Keim. Information visualization and visual data mining. *IEEE Trans. Vis. Comp. Graph.*, 8:1–8, January 2002.

[11] V. Kumar, J. Srivastava, and P.-N. Tan. Selecting the right interestingness measure for association patterns. In *Proc. ACM SIGKDD*, pages 32–41, 2002.