FpViz: A Visualizer for Frequent Pattern Mining

Carson Kai-Sang Leung[*]
Department of Computer Science
The University of Manitoba
Winnipeg, MB, Canada
kleung@cs.umanitoba.ca

Christopher L. Carmichael
Department of Computer Science
The University of Manitoba
Winnipeg, MB, Canada
umcarmi1@cs.umanitoba.ca

ABSTRACT

Over the past 15 years, numerous algorithms have been proposed for frequent pattern mining as it plays an essential role in many knowledge discovery and data mining (KDD) tasks. Most of these frequent pattern mining algorithms return the mined results in the form of textual lists containing frequent patterns showing those frequently occurring sets of items. It is well known that "a picture is worth a thousand words". The use of visual representation can enhance the user understanding of the inherent relations in a collection of frequent patterns. A few visualizers have been developed to visualize the input data or the mined results. However, most of these visualizers were not designed for visualizing the mined frequent patterns. In this paper, we develop a *visualizer for frequent pattern mining*. Such a visualizer—called *FpViz*—gives users an insight about the data, allows them to zoom in and zoom out, and provides details on demand. Moreover, FpViz is also equipped with several interactive features for effective visual support in the data analysis and KDD process for various real-life applications.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems—*human factors*; H.2.8 [**Database Management**]: Database Applications—*data mining*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces; I.4.0 [**Image Processing and Computer Vision**]: General—*image displays*

General Terms

Algorithms; Design; Experimentation; Human factors; Management; Measurement; Performance; Reliability

Keywords

Knowledge discovery and data mining, visual analytics, visual and interactive data analysis, visual support in the knowledge discovery process, data and knowledge visualization, frequent itemsets, visual data mining

1. INTRODUCTION

Frequent pattern mining [1, 20, 22, 23, 25, 26] aims to search for implicit, previously unknown, and potentially useful information in the form of *frequent patterns* (i.e., frequently occurring sets of items, which are also known as *frequent itemsets*). It plays an essential role in many knowledge discovery and data mining (KDD) tasks. Examples of these KDD tasks include the mining of association rules, correlation, sequences, episodes, maximal frequent patterns, and closed frequent patterns. Hence, frequent pattern mining is in demand in various real-life applications. Mined frequent patterns can answer many questions that help users make important decisions in various real-life situations. The following are some examples:

Q1. Store managers may want to find out how frequently certain kinds of vegetables (e.g., asparagus, broccoli) are purchased *individually* and how frequently are they purchased *together*? What kinds of vegetables are frequently purchased together with eggplants (e.g., {asparagus, broccoli, eggplants, peas})?

Q2. Botanists may want to discover which features or properties associated with edible mushroom are frequently observed?

Q3. University administrators may want to know which popular elective courses (e.g., {AI, Bioinformatics, Computational Geometry}) are frequently taken together by students?

Q4. Bookstore owners may want to know which books are also bought by customers who bought a particular KDD book so that they could bundle these books together for customer convenience?

Q5. Internet providers may want to figure out what Webpages are frequently browsed by Internet users in a single session?

Q6. Service planners may want to know why the demand of some combinations of services is dropping so that they could cancel those combinations and put the resources on other demanding combinations?

Q7. Web administrators may want to find out which collection of Webpages is frequently updated by users? Which groups of users frequent update the Webpages?

Q8. Travel agencies may want to discover where are the favourite spots and when are the popular time for travel?

[*]Corresponding author: C.K.-S. Leung.

VAKD '09, June 28, 2009, Paris, France.

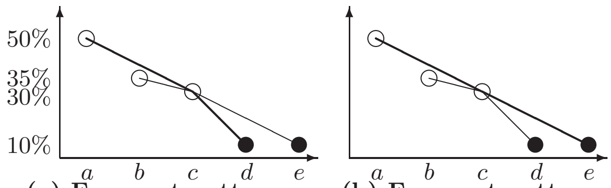
Q9. Phone service providers may want to find out where are the popular calling and receiving countries (e.g., {Canada, France}) for long-distance phone calls so that they could put these countries on their promotional package?

Q10. Security staff may want to know which parts of the building are frequently visited by employees or visitors?

To help answer the above questions in these real-life situations, numerous frequent pattern mining algorithms have been proposed over the past 15 years. However, most of the algorithms return a collection of frequent patterns in *textual form* (e.g., a very long unsorted list of frequent patterns). Consequently, users may not easily discover the knowledge and useful information that is embedded in the data.

Showing a collection of frequent patterns in *graphical form* can show the relations embedded in the data and help users understand the nature of the useful information and discovered knowledge. Hence, researchers have also considered visual analytics [8, 16, 17, 18, 21, 29, 32, 33, 37, 39] and visualization techniques [9, 13, 14, 34] to assist users in gaining insight into massive amounts of data or information. Visualization systems like Spotfire [2], VisDB [15] and Polaris [35] have been developed for visualizing data. For systems that visualize the mining results, the focus has been mainly on results such as clusters [19, 30], decision trees [3], social networks [4] or association rules [7]. However, *not* many visualizers were designed for visualizing frequent patterns.

Recently, some researchers have shown interests in visualizing frequent patterns. For example, Yang [38] developed a system that can visualize frequent patterns. However, his system was primarily designed to visualize association rules, and it does not scale very well in assisting users to immediately see certain useful information (such as exact frequencies or support) of a very large number of frequent patterns. As another example, Munzner et al. presented a visualizer called PowerSetViewer (PSV) [28], which provides users with guaranteed visibility of frequent patterns in the sense that the pixel representing a frequent pattern is guaranteed to be visible by highlighting such a pixel. However, multiple frequent patterns may be represented by the same pixel. As the third example, we previously proposed a visualization system—called FIsViz [26]—that aims to visualize frequent patterns. FIsViz represents each frequent pattern by a polyline in a two-dimensional space. The location of the polyline indicates the exact frequency of the pattern explicitly. As a result, FIsViz enables users to visualize the mined results (i.e., frequent patterns) for many real-life applications. However, in some other applications (especially, when the number of frequent patterns is huge), FIsViz may not scale very well. Users may require more effort to be able to clearly visualize frequent patterns. The problem is caused by the use of polylines for representing frequent patterns. To elaborate, the polylines can be bent and/or can cross over each other. This makes it difficult to distinguish one polyline (representing a frequent pattern) from another. For example, in Figure 1, how to distinguish the two frequent patterns $\{a, c, d\}$ & $\{b, c, e\}$ from another two patterns $\{a, c, e\}$ & $\{b, c, d\}$ if we did not use different thickness for the polylines?



(a) Frequent patterns $\{a, c, d\}$ & $\{b, c, e\}$ **(b) Frequent patterns $\{a, c, e\}$ & $\{b, c, d\}$**

Figure 1: Polylines showing $\{a, c, d\}$ & $\{b, c, e\}$ vs. polylines showing $\{a, c, e\}$ & $\{b, c, d\}$.

Hence, some natural questions to ask are: Can we design a scalable system that helps users visualize frequent patterns effectively? Can we have an alternative representation that minimizes the bend and crossover of polylines? In response to these questions, we explored an alternative representation [27], which uses two half-screens to visualize the discovered knowledge about frequent patterns: one half of the screen showing all frequent patterns and another half showing their frequencies.

In this paper, we propose a visualizer that uses a full screen for visualizing frequent patterns. The proposed visualizer enhances the KDD process by providing answers to some important business questions (e.g., Q1–Q10 above). The **key contribution** of our work is a novel interactive and scalable *frequent pattern visualizer*, called *FpViz*, which provides users with effective visual support in the data analysis and KDD process. Specifically, FpViz uses orthogonal graphs for visualizing frequent patterns. The visualizer provides users with clear and explicit depictions about frequent patterns that are embedded in the data of interest. Hence, FpViz enables users to infer (i) patterns at a glance and (ii) answers to many questions encountered in various real-life applications. It also provides interactive features for constrained mining and interactive mining. Moreover, with FpViz, users can (i) efficiently find closed or maximal itemsets and (ii) properly formulate association rules from the displayed frequent patterns.

This paper is organized as follows. Next section briefly describes related work and background. In Section 3, we introduce our FpViz and describe its design; in Section 4, we present interactive features provided by FpViz. Section 5 shows evaluation results. Finally, conclusions are presented in Section 6.

2. RELATED WORK AND BACKGROUND

Developing effective visualization systems for KDD has been the subject of many studies. This line of research can be subclassified into two general categories: (i) systems for visualizing raw data and (ii) those for visualizing data analysis or data mining results. Examples of systems in the first category include Spotfire [2], independence diagrams [5], VisDB [15], and Polaris [35]. These systems were built for visualizing data. Systems in the second category focus on *visualizing the mining results*, which include clusters [19], decision trees [3, 10], association rules [6, 12, 38], and frequent patterns [26, 27, 28, 38, 40]. Let us briefly discuss below some relevant systems for visualizing association rules or frequent patterns.

Yang [38] designed a system mainly to visualize association rules—but can also be used to visualize frequent patterns—in a two-dimensional space consisting of many vertical axes. In his system, all domain items are sorted according to their frequencies and are evenly distributed along each vertical axis. A frequent pattern consisting of k items (i.e., a k-itemset) is then represented by a curve that extends from one vertical axis to another connecting k such axes. The thickness of the curve indicates the frequency (or support) of such a frequent pattern. However, such a representation suffers from the following problems: (i) The use of thickness only shows *relative* (but not *exact*) frequency of the patterns. Comparing the thickness of curves is not easy. (ii) Since items are sorted and *evenly* distributed along the axes, users only know some items are more frequent than the others, but cannot get a sense of how these items are related to each other in terms of their exact frequencies (e.g., whether item a is twice as frequent as, or just slightly more frequent than, item b). (iii) Although Yang's system is able to show both association rules and frequent itemsets, his system does not provide users with many interactive features, which are necessary if a large graph containing many items to be displayed.

Frequent itemset visualizer (FIsViz) [26] is one of the recently developed visualizers. It was designed for visualizing frequent itemsets. It represents a k-itemset represented by a polyline that connects k nodes (where each node represents an item in the k-itemset) in a two-dimensional space. The frequency (or support) of the i-th prefix of an itemset X is indicated by the position of the i-th node in the polyline representing X. For example, when $X = \{a, c, d\}$, the frequencies of its prefixes $\{a\}$ and $\{a, c\}$ are respectively indicated by the positions of nodes a and c in the polyline. Similarly, the frequency of the itemset $X = \{a, c, d\}$ is represented by the y-position of the node d in that polyline. See Figure 1(a). With such itemset representation, slopes of different sectors of a polyline can vary. In other words, the entire polyline may not be a straight one (i.e., it may be bent). Moreover, polylines representing different itemsets may cross each others. This makes it difficult for users to distinguish one sector of a polyline from another. See Figure 1.

3. FpViz: OUR PROPOSED FREQUENT PATTERN VISUALIZER

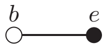
In this section, we present our proposed **F***requent* **p***attern* **Vi***suali***z***er* (**FpViz**). Here, FpViz is connected to a frequent pattern mining algorithm (e.g., FP-growth [11]), which finds frequent patterns from transaction database. Once frequent patterns are found, FpViz effectively displays them for the data analysis. Note that FpViz is not confined to using FP-growth for frequent pattern mining. It can use some other frequent pattern mining algorithms (e.g., , DCF [20] for constrained mining, UF-streaming [23] for stream mining, UF-growth [24] for uncertain data mining, Apriori [1]).

Like FIsViz [26], our proposed FpViz also shows frequent patterns consisting of k items (i.e., k-itemsets) in a two-dimensional space. The x-axis shows the n domain items. We allow the user to specify his preference on the ordering of these domain items. For example, the user can arrange the items in (i) non-ascending frequency order, (ii) lexicographical order, or (iii) some other orders (e.g., put those items of interest—such as promotional items—on the left and less interesting items on the right side of the x-axis) for constrained mining. The y-axis shows the frequencies of the frequent patterns.

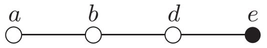
Unlike FIsViz (which represents frequent patterns as polylines), the basic representation for our proposed FpViz is an orthogonally laid out node-link diagram. According to graph aesthetics [31, 36], reducing the number of edge crossings can improve the legibility of graphs. Similarly, assigning uniform lengths to edges and minimizing bends can enhance the legibility of the node-link diagram. Since our datasets are potentially very large, a primary criterion in our design is to minimize edge crossings and bends. We, therefore, adopted an orthogonal layout mechanism that preserves edge crossings to a minimum. Bends occur only at 0° or 90° angles. As a result, FpViz minimizes crossings, facilitating legibility and visual comprehension.

3.1 Representing Frequent Patterns

Our proposed FpViz represents each frequent pattern X consisting of k items (i.e., k-itemset) by a horizontal line connecting k nodes (represented by k circles), where each node represents an item within the frequent pattern X. For example, the 2-itemset $\{b, e\}$ is represented by a horizontal line connecting two circles (where each circle represents an item), as follows:



Note that, between the two circles, one of them is filled. The filled circle (i.e., disc) represents the last item (according to the item order $\mathcal{R}$) in the frequent pattern $\{b, e\}$. As another example, the 4-itemset $\{a, b, d, e\}$ is represented by a horizontal line connecting four circles (where the last one is filled), as follows:



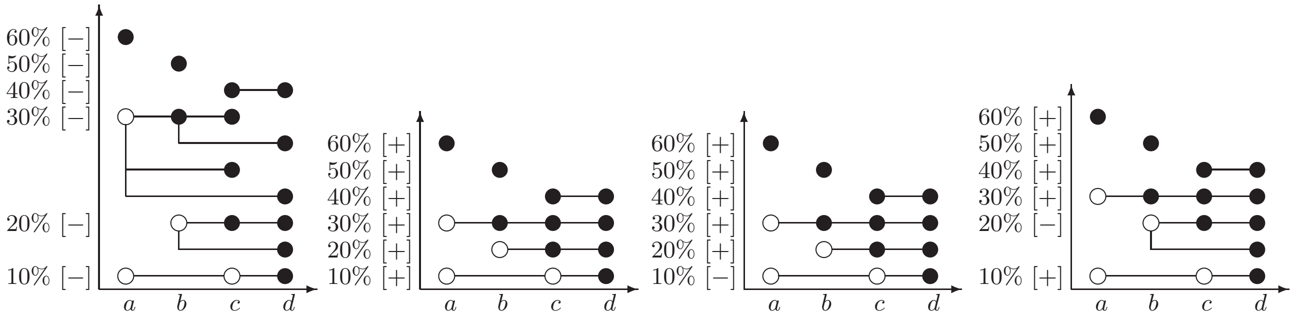
For singletons (i.e., 1-itemsets), they are represented by just filled circles (or filled diamonds for user convenience) in FpViz. For example, the singleton $\{e\}$ is represented as:

To summarize, each frequent pattern consisting of k items (i.e., k-itemset) is represented by a *horizontal line* connecting k circles, with the last circle filled.

3.2 Showing the Frequencies of Frequent Patterns

The frequency of a frequent pattern consisting of k items (which is represented by a horizontal line connecting k circles with the last circle filled) is indicated by the y-value (i.e., y-position) of the filled circle. This way of showing the frequencies work reasonable well when each frequent pattern has a distinct frequency (i.e., at most one horizontal line for each frequency value—the y-value).

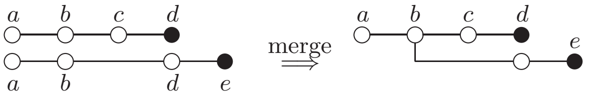
However, for many real-life applications, it is not uncommon that multiple frequent patterns happen to have the same frequency. In these situations, we apply *compression*



(a) The expanded view ("expand all") **(b) The collapsed view (the default view)** **(c) An expanded view ("expand the 10% line")** **(d) An expanded view ("expand the 20% line")**

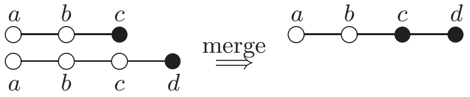
Figure 2: Expanded and collapsed views for visualizing frequent patterns with our proposed FpViz.

techniques to our proposed FpViz: If the two frequent patterns X and Y of the same frequency share the same prefix, then their common prefix is merged. The suffixes of X and Y are then branching out from the last item of the common prefix. For example, if frequent patterns $\{a, b, c, d\}$ and $\{a, b, d, e\}$ (which share the same prefix $\{a, b\}$) are of the same frequency, they can be represented as follows:



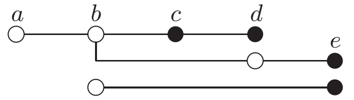
Here, $\{c, d\}$ and $\{d, e\}$ are two branches of the common prefix $\{a, b\}$.

A special case of the merge occurs when a suffix of Y is branching out from the last item of X (i.e., X is a prefix of Y). In this case, the two horizontal lines representing the two frequent patterns X and Y would be merged into one line. For example, for frequent patterns $\{a, b, c\}$ and $\{a, b, c, d\}$, the former is a prefix of the latter. Hence, these two frequent patterns can be merged to form the following:

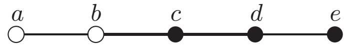


Here, the filled circle d indicates the last item of the frequent pattern $\{a, b, c, d\}$, whereas the filled circle c indicates the last item of the prefix $\{a, b, c\}$. Note that this merge helps reduce the number of horizontal lines to be drawn (i.e., reduce the amount of vertical space required for displaying the frequencies of all the frequent patterns).

When the number of mined frequent patterns is not huge, the merging of patterns with their prefixes having the same frequencies (e.g., the case for $\{a, b, c\}$ and $\{a, b, c, d\}$) reduces the amount of vertical space required. However, when the number of mined frequent patterns is huge, we may still run out of vertical space to fit all horizontal lines representing all the mined frequent patterns—even when merging is applied. Hence, we need to apply further compression technique as follows. To reduce the amount of space required in the y-direction, if multiple frequent patterns (say, m frequent patterns represented by m' horizontal lines, where $m' \leq m$) have the same frequency, they are projected or collapsed into *one horizontal line* (instead of m' lines). For instance, frequent patterns $\{a, b, c\}$, $\{a, b, c, d\}$, $\{a, b, d, e\}$ and $\{b, e\}$ are of the same frequency:



These $m = 4$ frequent patterns (represented by $m' = 3$ horizontal lines) are collapsed into one horizontal line, as shown below:



By so doing, each existing frequency value would be represented by one—and only one—horizontal line. For example, Figure 2(a) shows $m = 13$ frequent patterns represented by two disjointed filled circles for singletons $\{a\}$ & $\{b\}$ and $m' = 8$ horizontal lines for other 11 non-singleton frequent patterns. Figure 2(b) shows how these $m' = 8$ horizontal lines are collapsed into four lines by using our proposed FpViz. The resulting view shows two disjointed filled circles and four lines, which represent $m = 13$ frequent patterns having $2 + 4 = 6$ distinct frequencies.

It is important to note that (though it may not be obvious in this black-and-white version of our paper), FpViz uses the color of the circle to indicate the number of occurrences of an item within those frequent patterns of the same frequency. For example, a lighter circle a indicates that a only occurs in one frequent pattern, whereas a darker circle b indicates that b occurs more often (e.g., in four frequent patterns $\{a, b, c\}, \{a, b, c, d\}, \{a, b, d, e\}$ and $\{b, e\}$). Moreover, the thickness of the line indicates the number of horizontal lines that were collapsed into one.

3.3 Collapsing and Expanding the Horizontal Lines

Our proposed FpViz normally shows frequent patterns in the (default) *collapsed view* so as to reduce the mount vertical space required for displaying all patterns. As this collapsed view may hide some details, FpViz provides the user with the option to expand on any portion of the graph that is interesting to him by clicking the [+] button. By so doing, the user would be able to clearly get all the details.

As an example, when the user clicks the [+] button for frequency=10%, FpViz expands the horizontal line representing frequent patterns of frequency=10%. Consequently, the user obtains the expanded view as presented in Figure 2(c), which shows that such a horizontal line represents the frequent pattern $\{a, c, d\}$. Similarly, when the user clicks the [+] button for frequency=20%, FpViz expands the horizontal line representing frequent patterns of frequency=20%. The user then obtains the expanded view as presented in Figure 2(d), which shows that such a horizontal line represents three frequent patterns $\{b, c\}$, $\{b, c, d\}$ and $\{b, d\}$. Note that the user is not confined to clicking only one [+] button, he could click all six [+] buttons to obtain an expanded view as shown in Figure 2(a).

3.4 Observations

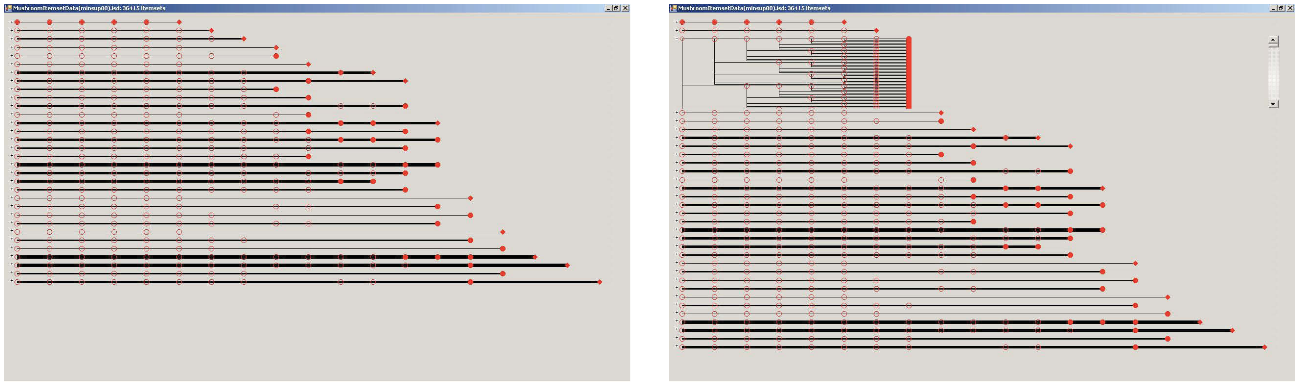
With this representation of frequent patterns and their frequencies in our proposed FpViz, users can observe the following from the default collapsed view:

1. By default, FpViz arranges the domain items in non-ascending frequency order. As a result, the most frequently occurring item (which with the highest frequency) appears on the left side and the least frequently occurring one appears on the right side. In other words, users can easily get an insight about the frequency ranking of all the domain items by walking along the x-axis. For example, we observed from Figure 2(b) that item a is the most frequent domain item, which is followed by items b and c, and item d is the least frequent domain item. (It is important to note that, users are not confined to this ordering; they can choose other ordering $\mathcal{R}$ to arrange all items in the domain.)
2. FpViz gives information about frequency distribution of items in the frequent patterns. For example, users can tell how many distinct frequency values can an item take on (in any frequent pattern) by counting the number of its y-values. For example, we observed from Figure 2(b) that item b takes on tree distinct frequency values—namely, 20%, 30% and 50%—by counting the number of circles for b.
3. The frequency of any subset of a frequent pattern X is guaranteed to be higher than or equal to that of X. Hence, the disjointed filled circle (or diamond) representing any singleton subset of X (or the horizontal line representing any non-singleton subset of X) is guaranteed to appear on or above the horizontal line representing X. For example, let us consider $\{c, d\}$, which is a subset of frequent pattern $\{a, c, d\}$. We observed from Figure 2(b) that their frequencies are 40% and 10%, respectively. In other words, the frequency of the subset ($\{c, d\}$) is higher than that of the frequent pattern $\{a, c, d\}$. Similarly, we observed that the frequency of both $\{b, c\}$ and $\{b, c, d\}$ are the same (of 20%).
4. Conversely, the frequency of any superset of a frequent pattern X is guaranteed to be lower than or equal to that of X. Hence, the horizontal line representing any superset of X is guaranteed to appear on or below the horizontal line representing X. For example, we observed from Figure 2(b) that the frequency of the superset ($\{a, c, d\}$) is lower than that of the frequent pattern $\{c, d\}$. Similarly, we observed that the frequency of both $\{b, c\}$ and $\{b, c, d\}$ are the same.
5. The highest frequency value for each item u is the frequency of the singleton $\{u\}$. For example, the highest frequency value of item b is 50%, which the frequency of the singleton $\{b\}$.
6. If a horizontal line starts with a hollow circle and follows by a filled circle, then users can reveal without requiring any expansion (i.e., without clicking any [+] button) and guarantee that a frequent pattern consisting of only two items exists with that frequency. For example, we observed from Figure 2(b) that a horizontal line at frequency=30% starts with a hollow circle for item a and follows by a filled circle for item b. We then knew, without expanding such a horizontal line, that there exists frequent pattern $\{a, b\}$ and its frequency is 30%. (This observation could be confirmed by clicking the [+] button for the 30% line.)
7. If a horizontal line involves *only two* items, users do not need to expand such a line (i.e., do not need to click the [+] button) to get the complete information about the frequent pattern consisting of only two items. The frequency of this pattern is clearly indicated by the y-position of the last item. Moreover, if such a horizontal line starts with a diamond, then this line gives additional information that the 2-itemset and its singleton prefix are of the same frequency. For example, we observed from Figure 2(b) that there exists a frequent pattern $\{c, d\}$ consisting of only two items and its frequency is 40%. Moreover, its singleton prefix $\{c\}$ is also of frequency 40%.
8. For a horizontal line representing frequency=y%, if the first circle (representing item u) is filled, then the singleton has frequency of y%. For example, the frequency of singleton $\{b\}$ is 50% and that of $\{c\}$ is 40%.
9. For a horizontal line representing frequency=y%, if the first circle (filled or hollow) represents an item u and the second circle (representing item v) is filled, then the frequency of the frequent pattern $\{u, v\} = y$%. For example, the frequencies of $\{c, d\}$ and $\{a, b\}$ are 40% and 30%, respectively.
10. For a horizontal line that involves more than two items and does not have a filled circle in its second position in the collapsed view, FpViz provides information about the *absence* of any items from frequent patterns of that frequency. For example, since item b does not appear in the 10% line, b is guaranteed not to appear in any patterns of that frequency.

    However, one may not be able to easily determine the contents of the frequent patterns represented by this line. For example, the 10% line in Figure 2(b) could represent (i) frequent patterns $\{a, d\}$ and $\{c, d\}$ and/or (ii) frequent pattern $\{a, c, d\}$. (This explains why FpViz provides users with [+] buttons for expanding the lines to clearly obtain detailed information.)

In addition, users can also observe the following when they click one or more [+] buttons to expand horizontal lines of their interest:

11. After expanding a horizontal line, users can obtain the cardinality k of the k-itemset (which is represented by


MushroomItemsetData(minsup80).isd: 36415 itemsets
MushroomItemsetData(minsup80).isd: 36415 itemsets

(a) The collapsed view (the default view)

(b) An expanded view (when one horizontal line is expanded)

Figure 3: Snapshots of our proposed FpViz showing the collapsed and expanded views.

each expanded line) by counting the number of circles on such an expanded line (from the leftmost circle to a filled one). For example, the cardinality of $\{b, c\}$ is 2 because there are 2 circles from the leftmost one to the first filled one along the 20% line shown in Figure 2(d). Similarly, the cardinality of $\{b, c, d\}$ is 3 because there are 3 circles from the leftmost one to the second filled one). While users can count the number of circles to determine the cardinality of a frequent pattern, FpViz provides the feature called *query on cardinality* for user convenience. See Section 4.

12. The first portion of an expanded horizontal line represents a prefix of a frequent pattern X. If such a portion does not end with a filled circle, then the frequency of such a prefix of X is different from that of X. For example, we observed from Figure 2(c) that the frequent pattern $\{a, c, d\}$ does not have the same frequency as its prefix $\{a, c\}$ (10% vs. 30%) because the circle representing c in $\{a, c\}$ is hollow.

13. If there are more than one filled circle on an expanded horizontal line, then a frequent pattern X and at least one of its prefix have the same frequency. For example, as shown in Figure 2(d), frequent patterns $\{b, c, d\}$ and its prefix $\{b, c\}$ have the same frequency of 20%.

14. Once users observe $sup(X)$ and $sup(X \cup Y)$, they can compute the support, confidence and lift of association rule $X \Rightarrow Y$ using $sup(X \cup Y)$, $\frac{sup(X \cup Y)}{sup(X)}$ and $\frac{sup(X \cup Y)}{sup(X) \times sup(Y)}$ respectively. Moreover, if $sup(X) = sup(X \cup Y)$, then users can easily determine that $conf(X \Rightarrow Y) = 100\%$.

3.5 Discussions

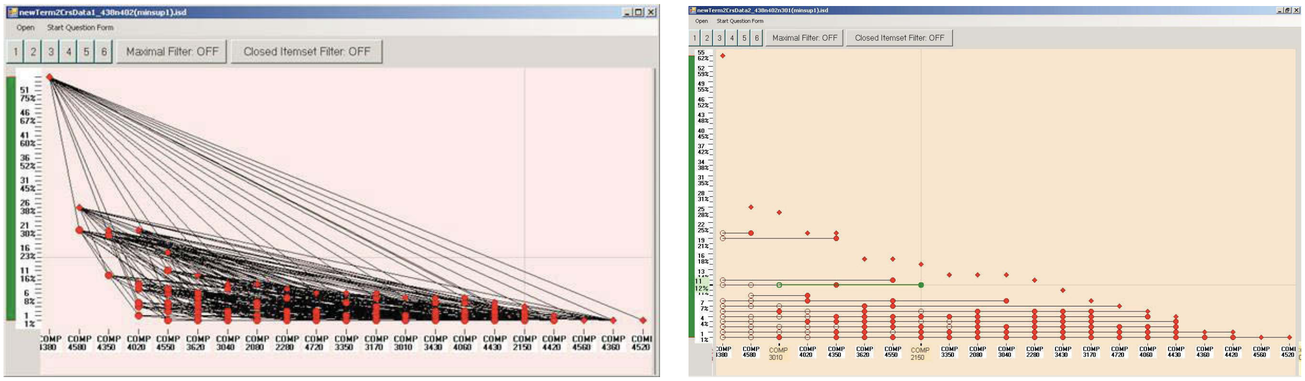
How to represent frequent patterns with a huge number of distinct frequency values? Recall that FpViz applies the compression technique for merging several horizontal lines that represent frequent patterns of the same frequency into one line. To a further extent, if the number of distinct frequencies in the mined results exceeds the number of vertical pixels (or the number of horizontal lines allocated for the space), we can apply the compression technique further. To elaborate, we not only can compress the frequent patterns of the same frequency (e.g., sup=72%) but can also compress those frequent patterns with the same frequency range (e.g., $sup \in [70\%, 75\%]$). The lesser the allocated space, the broader would be the frequency range to be used in compression.

Besides compressing/merging several horizontal lines (which represent several frequent patterns) into a single line, FpViz can also provide users with some alternative options. For example, we can allow users to pick an option to display and visualize only those *closed frequent patterns* or *maximal frequent patterns* (instead of showing all frequent patterns). A frequent pattern X is *closed* if there does not exist any proper superset of X having the *same frequency as* X, and a frequent pattern Y is *maximal* if there does not exist any proper superset of Y that is also *frequent*. By visualizing only closed or maximal frequent patterns, FpViz can reduce the number of horizontal lines that need to be shown.

How to represent frequent patterns comprising a large number of domain items? While the above dealt with the scalability issue for the y-direction (i.e., scalable for large range of frequencies), we discuss here the scalability issue for the x-direction. More specifically, what if the number of items exceeds the number of pixels allocated for the x-direction? The challenge is that the y-direction contains frequency information (which are numerical data) but the x-direction contains items (which are categorical data). In FpViz, we can apply the following two techniques. First, if items are arranged in decreasing frequency order, then we can group several adjacent items together into a "mega"-item. For example, "mega"-item$_1$ represents all the domain items having the top 10 frequencies, "mega"-item$_2$ represents all the domain items having the next 10 highest frequencies, etc. Alternatively, we can group items according to some hierarchy or taxonomy. For example, we could group items "apples", "bananas" & "cherries" into a mega-item "fruits". Similarly, we could also group "donuts" & "egg-tarts" into a mega-item "snack".

4. INTERACTIVE FEATURES OF FpViz

The above representation of FpViz allows users to get an insight about the overview distribution of raw data and the analysis results (in the default collapsed view). It also provides users with some relevant details (in the expanded view). See Figure 3 for snapshots (of these two views). In



(a) FIsViz [26]

(b) Our proposed FpViz

Figure 4: Two visualizers showing the same set of frequent patterns mined from a student-course DB.

this section, we describe some additional interactive features of FpViz. While these features are not essential, they provide user convenience.

Query on frequency. With FpViz, users can easily find all *frequent items* and/or *frequent patterns* (i.e., with frequencies exceeding the user-specified minimum frequency threshold *minsup*) by ignoring everything that lies below the "threshold line" y=*minsup* (i.e., ignoring the lower portion of the graph). To a further extent, the representation of frequent patterns in FpViz leads to effective *interactive mining*. To elaborate, with FpViz, users can see what (and how many) frequent patterns are above a certain frequency. Based on this information, users can freely adjust *minsup* by moving the slider (see the green bar on the left side in Figure 4(b))—which controls *minsup*—up and down along the y-axis to find an appropriate value for *minsup*. Moreover, FpViz also provides two related features:

(i) It allows users to interactively adjust *minsup* and automatically counts the number of patterns that satisfy *minsup*. By doing so, users can easily find top-N frequent patterns.

(ii) It also allows users to pose a range query on frequency (by specifying both minimum and maximum frequency thresholds *minsup* and *maxsup*) and then shows all patterns with frequencies falling within the range [*minsup*, *maxsup*].

Query on cardinality. FpViz allows the user to pose a query on cardinality, and it only shows frequent patterns of the user-specified cardinality k. Moreover, FpViz also allows users to pose a range query on cardinality so that only those frequent patterns with cardinality k within the user-specified range [k_{min}, k_{max}] are drawn.

Query on frequent patterns. FpViz also allows users to interactively select certain items of interest (e.g., promotional items in a store) and to pose queries on frequent patterns. Examples of these queries include the following: (i) "Find all frequent patterns containing *some* of selected items"; (ii) "Find all frequent patterns containing at least *all* of the selected items"; and (iii) "Find all frequent patterns *not* containing any of the selected items".

Query on relationships among frequent patterns. Recall that users can easily find the prefixes or extensions of a frequent pattern that share the same frequencies. However, sometimes prefixes or extensions (or more general case, subsets or supersets) of a frequent pattern X may have different frequencies than that of X. Hence, FpViz provides interactive features to highlight these subsets or supersets (prefixes or extensions) of a pattern of user interests.

Details-on-demand. Details-on-demand consists of techniques that provide more details whenever the user requests them. The key idea is that FpViz gives users an overview of the entire dataset and then allows users to interactively select parts of the overview for which they request more details—by hovering the mouse over different parts of the display. Specifically, FpViz supports details-on-demand in the following ways:

(i) When *the mouse hovers on a segment of a horizontal line* connecting two nodes (say, representing frequent patterns x and y), FpViz shows a list of frequent patterns containing both x and y. Selecting a frequent pattern in the list instantly highlights the specific segment it is contained in, as well as both of its connecting nodes, so that users can see where the segment starts and ends.

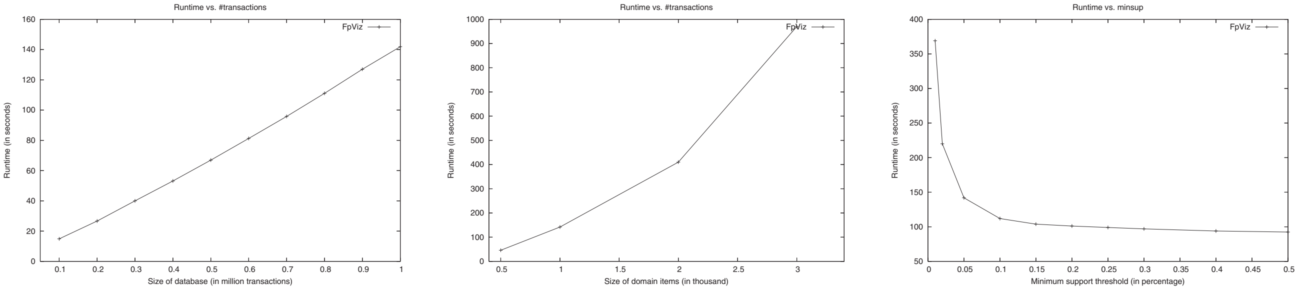
(ii) When the *mouse hovers over a node*, FpViz shows a list of all frequent patterns contained in all the line segments starting or ending at this node. Selecting a frequent pattern from the list instantly highlights the line it is contained in.

(iii) When the *mouse hovers over a pixel* in the display (even if it is not part of the graph), a small box appears showing the frequency and frequent patterns encoded by the mouse position. This is particularly useful when users need to see among the vast array of line segments what a particular point in the display refers to.

Guaranteed visibility. Our proposed FpViz allows users to specify his preference on visualization of frequent patterns. For example, if users are interested in finding those patterns containing fruits, FpViz ensures that all corresponding horizontal lines are clearly visible.

5. EVALUATION RESULTS

In this section, we show our results on evaluating the proposed FpViz. Here, we conducted four sets of evaluation



(a) Runtime vs. sizeof(*TDB*) **(b) Runtime vs. #domain items** **(c) Runtime vs. *minsup***

Figure 5: Performance of FpViz.

tests. In the first set, we tested functionality of our FpViz by showing how it can be applicable in various scenarios or real-life applications. In the second set, we tested performance of our FpViz. In the last two sets, we evaluated readability and interactivity of FpViz.

5.1 Evaluating the Functionality of FpViz

In the first set of evaluation tests, we compared our proposed FpViz with FIsViz [26]. We considered many different real-life scenarios. For each scenario, we determined whether these systems can handle the scenarios. If so, we examined how these systems display the mined results. The evaluation results show that our FpViz was as effective as FIsViz in all these scenarios. A few samples of these scenarios are shown below.

Q1(a) What kinds of vegetables are frequently purchased by customers? Frequently purchased vegetables are patterns with high frequency. With FIsViz, as polylines representing itemsets may cross each other, users may *not* be able to easily see the itemsets of high frequency if they are in the dense or clustered area of the display. In contrast, our FpViz shows all frequent patterns by horizontal lines, which are easily visible and never cross any other horizontal lines. Let us compare Figure 4(a) with Figure 4(b). The former was a snapshot of FIsViz and the latter was a snapshot of FpViz. These two snapshots both show the same set of frequent patterns.

Q1(b) How frequently are these vegetables purchased individually and how frequently are purchased together? Depending on the density of the display in FIsViz, the frequencies of some itemsets may *not* be too visible if they are in the dense or clustered areas of the display. In contrast, users can easily obtain the frequencies of patterns from our FpViz because there is no line crossing.

Q1(c) What kinds of vegetables that are frequently purchased together with eggplants? Both FIsViz and FpViz provide users with a feature of handling queries on frequent patterns containing some specific items (in this scenario, eggplants).

We observed from all scenarios (including the above three samples) that our proposed FpViz either retained the existing features of FIsViz (e.g., for Q1(c)) or provided additional improvements over FIsViz (e.g., for Q1(a) & (b)).

5.2 Evaluating the Performance of FpViz

In the performance test, we used (i) several IBM synthetic datasets [1], (ii) some real-life databases (e.g., mushroom dataset) from UC Irvine Machine Learning Depository, (iii) some CNN documents, and (iv) a student-course database for our university. The results produced are consistent. In the first experiment, we varied the size of the database *TDB*. We measured the time both for mining frequent patterns (by using FP-growth [11]) and for constructing the display layout (by using our proposed FpViz). The results showed that the runtime (which includes CPU and I/Os) increased linearly with the number of transactions in the database. See Figure 5(a). In the second experiment, we varied the number of items in the domain. The results showed that the runtime increased when the number of domain items increased. See Figure 5(b). In the third experiment, we varied the user-defined frequency threshold. When the threshold increased, the number of patterns that satisfy the threshold (i.e., frequent patterns to be displayed) decreased, which in turn led to a decrease in runtime. See Figure 5(c).

5.3 Evaluating the Readability of the Mined Results Shown by FpViz

To assess the effectiveness of conveying frequent pattern relationships, we carried out a user evaluation with FpViz. The evaluation was primarily case-based, within which several types of users were required to solve many different questions based on the visualizations of a given dataset (e.g., a database containing information about courses taken by students (see Figure 4(b)). Therefore, the scenario was that users need to identify a set of relationships and make decisions based on their observations.

We recruited 24 participants and separated them into two groups: (i) those who have data mining background and (ii) those who do not. None of the participants (regardless which of the two groups) was exposed to any form of visualization for frequent patterns—including our proposed FpViz.

To test the expressiveness of our visualization, we formulated two types of questions: multiple choices and those open-end ones that require participants to perform some level of analytical reasoning with the visualization. Sample questions include the following:

1. Which course is most frequently taken (i.e., course with highest enrolment)?
2. Which course is the next/second most frequently taken (i.e., course with second highest enrolment)?
3. Which two courses are most frequently taken?
4. What course is least frequently taken (i.e., course with lowest enrolment)?
5. What three courses are taken together by exactly four students?
6. How many students are taking COMP 4350 together with 4380?
7. What three courses are taken together by four students?
8. How would you use this chart to make any changes in terms of how 3rd and 4th year courses are offered and/or distributed?
9. If you were to reduce the offerings of 4th year courses, which ones would you select? and why (i.e., how did the diagrams lead to your conclusions)?
10. If you were to schedule exams of these courses, which pair of courses would you avoid scheduling on the same day?

We began the evaluation by presenting our FpViz and asking the participants to explore it at their own will. We did not give them any information regarding what the symbols and representations meant in the visualization. We first questioned them on what they were able to identify. Evaluation results showed that all the participants were able to identify the basic meaning behind the representations (e.g., that frequency was assigned to the y-axis, and courses to the x-axis). Participants were also able to identify the most frequently taken courses, without having us to tell them the answer.

Afterwards, we gave the participants detailed information on how to read the graphs and what the various lines and circles meant. This information was then followed by a set of questions that queried into the participants' ability to simply read the graph, with a set of close-ended questions. Evaluation results showed that a majority of the participants were able to correctly answer most—and some even correctly answer all—of the questions. Statistically, the average accuracy rate was above 83% and three participants obtained an accuracy rate of 100%.

Several interesting observations were found at the post-evaluation interview. For example, the participants told us that, while we gave the participants a multiple choice list to help them identify the answers, they did not use the multiple choice questions to guide their selection. Instead, their curiosity in understanding the visualizations led them to answer the questions by looking at the graphs first, and then confirming their answer with one of the choices provided. Moreover, the results were similar in both participant groups (with or without data mining background). This shows that the easy readability of FpViz. This also suggested that, with very little training, participants felt comfortable to use the visualizations. Furthermore, they were able to quickly assimilate the representations, to the point at which they were able to answer all questions adequately.

Finally, we asked the participants a set of open-ended questions. Each participant completed the evaluation separately (i.e., no discussion among the participants). The results showed that the participants were able to make the best use of visualization for answering these questions.

5.4 Evaluating the Benefits of the Interactive Features Provided by FpViz

We also evaluated the effectiveness of the interactivity of our FpViz on various datasets mentioned above. We divided a set of 24 participants into two groups. The first group performed some tasks using *only* the essential features (i.e., without using any interactive feature described in Section 4), and then used both the essential and the interactive features to answer some similar but not identical tasks. The second group did so in the reverse order (i.e., with interactive features and then without interactive features). By so doing, we avoid measuring the unwanted effect of learning (i.e., participants may learn from the first set of tasks). We observed the following from the results: (i) The participants were able to correctly answer all the questions using the interactive features. (ii) Most participants were able to do so without using the interactive features, but required much longer time. On average, participants took about 5 minutes to complete all the questions when using the interactive features, but took longer than 12 minutes to complete all the questions when not using any interactive feature. This indicated that, while interactive features are not essential, they provide convenience to users (and saved their time). Hence, it is beneficial to use the interactive features provided by FpViz.

6. CONCLUSIONS

Many frequent pattern mining algorithms return a collection of the data analysis results in the form of a textual list of frequent patterns. This list can be very long and difficult to comprehend. Since "a picture is worth a thousand words", it is desirable to have visualization systems. However, many existing visualization systems were not designed to show frequent patterns. To improve this situation, we proposed and developed a powerful *frequent pattern visualizer* called *FpViz*, which provides users with explicit and easily-visible information among the frequent patterns. Specifically, FpViz represents frequent patterns as horizontal lines in a two-dimensional graph. If multiple patterns have the same frequency, the corresponding lines representing these patterns are collapsed into one line. With this compression technique, FpViz allows the user to expand some portion (or all) of the "collapsed" mining results for data or result exploration. Moreover, FpViz also provides users with additional nice interactive features. Evaluation results showed the effectiveness of FpViz in terms of functionality, performance, readability, and interactivity. FpViz provides users with visual support for visual analytics as well as knowledge discovery and data mining (KDD).

7. ACKNOWLEDGMENTS

This project is partially supported by Natural Sciences and Engineering Research Council of Canada (NSERC) in the form of research grants.

8. REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proc. VLDB 1994*, pp. 487–499.

[2] C. Ahlberg. Spotfire: an information exploration environment. *SIGMOD Record*, **25**(4), pp. 25–29, 1996.

[3] M. Ankerst et al. Visual classification: an interactive approach to decision tree construction. In *Proc. KDD 1999*, pp. 392–396.

[4] P. Appan et al. Summarization and visualization of communication patterns in a large-scale social network. In *Proc. PAKDD 2006*, pp. 371–379.

[5] S. Berchtold et al. Independence diagrams: a technique for visual data mining. In *Proc. KDD 1998*, pp. 139–143.

[6] J. Blanchard et al. Interactive visual exploration of association rules with rule-focusing methodology. *KAIS*, **13**(1), pp. 43–75, 2007.

[7] C. Brunk et al. MineSet: an integrated system for data mining. In *Proc. KDD 1997*, pp. 135–138.

[8] S.-M. Chan et al. Maintaining interactivity while exploring massive time series. In *Proc. IEEE VAST 2008*, pp. 59–66.

[9] C.H. Chih and D.S. Parker. The persuasive phase of visualization. In *Proc. KDD 2008*, pp. 884–892.

[10] J. Han and N. Cercone. RuleViz: a model for visualizing knowledge discovery process. In *Proc. KDD 2000*, pp. 244–253.

[11] J. Han et al. Mining frequent patterns without candidate generation: a frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, **8**(1), pp. 53–87, 2004.

[12] H. Hofmann et al. Visualizing association rules with interactive mosaic plots. In *Proc. KDD 2000*, pp. 227-235.

[13] T. Iwata et al. Probabilistic latent semantic visualization: topic model for visualizing documents. In *Proc. KDD 2008*, pp. 363–371.

[14] D.A. Keim. Information visualization and visual data mining. *IEEE TVCG*, **8**(1), pp. 1–8, 2002.

[15] D.A. Keim and H.-P. Kriegel. Visualization techniques for mining large databases: a comparison. *IEEE TKDE*, **8**(6), pp. 923–938, 1996.

[16] D.A. Keim and D. Oelke. Literature fingerprinting: a new method for visual literary analysis. In *Proc. IEEE VAST 2007*, pp. 115–122.

[17] D.A. Keim and J. Schneidewind (eds.). Special issue on visual analytics. *SIGKDD Explorations*, **9**(2), 2007.

[18] D.A. Keim et al. Monitoring network traffic with radial traffic analyzer. In *Proc. IEEE VAST 2006*, pp. 123–128.

[19] Y. Koren and D. Harel. A two-way visualization method for clustered data. In *Proc. KDD 2003*, pp. 589–594.

[20] L.V.S. Lakshmanan, C.K.-S. Leung, and R.T. Ng. Efficient dynamic mining of constrained frequent sets. *ACM TODS*, **28**(4), pp. 337–389, 2003.

[21] H. Lam et al. Session viewer: visual exploratory analysis of web session logs. In *Proc. IEEE VAST 2007*, pp. 147–154.

[22] C. K.-S. Leung. Frequent itemset mining with constraints. To appear in *Encyclopedia of Database Systems*, Springer, 2009.

[23] C.K.-S. Leung and B. Hao. Mining of frequent itemsets from streams of uncertain data. In *Proc. IEEE ICDE 2009*, pp. 1663–1670.

[24] C.K.-S. Leung et al. A tree-based approach for frequent pattern mining from uncertain data. In *Proc. PAKDD 2008*, pp. 653–661.

[25] C.K.-S. Leung et al. CanTree: a tree structure for efficient incremental mining of frequent patterns. In *Proc. IEEE ICDM 2005*, pp. 274–281

[26] C.K.-S. Leung et al. FIsViz: a frequent itemset visualizer. In *Proc. PAKDD 2008*, pp. 644–652.

[27] C.K.-S. Leung et al. WiFIsViz: effective visualization of frequent itemsets. In *Proc. IEEE ICDM 2008*, pp. 875–880.

[28] T. Munzner et al. Visual mining of power sets with large alphabets. Technical report UBC CS TR-2005-25, Dept. of Computer Science, UBC, Canada, 2005.

[29] D. Oelke et al. Visual evaluation of text features for document summarization and analysis. In *Proc. IEEE VAST 2008*, pp. 75–82.

[30] G. Pölzlbauer et al. A vector field visualization technique for self-organizing maps. In *Proc. PAKDD 2005*, pp. 399–409.

[31] H.C. Purchase et al. Validating graph drawing aesthetics. In *Proc. GD 1995*, pp. 435–446.

[32] J. Scholtz. Beyond usability: evaluation aspects of visual analytic environments. In *Proc. IEEE VAST 2006*, pp. 145–150.

[33] T. Schreck et al. Visual cluster analysis of trajectory data with interactive Kohonen Maps. In *Proc. IEEE VAST 2008*, pp. 3–10.

[34] R. Spence. *Information Visualization: Design for Interaction - 2e.* Prentice Hall, 2007.

[35] C. Stolte et al. Query, analysis, and visualization of hierarchically structured data using Polaris. In *Proc. KDD 2002*, pp. 112–122.

[36] C. Ware et al. Cognitive measurements of graph aesthetics. *Information Visualization*, **1**(2), pp. 103-110, 2002.

[37] P.C. Wong and J. Thomas. Visual analytics. *IEEE CG&A*, **24**(5), pp. 20–21, 2004.

[38] L. Yang. Pruning and visualizing generalized association rules in parallel coordinates. *IEEE TKDE*, **17**(1), pp. 60–70, 2005.

[39] X. Yang et al. A visual-analytic toolkit for dynamic interaction graphs. In *Proc. KDD 2008*, pp. 1016–1024.

[40] J. Yuan et al. From frequent itemsets to semantically meaningful visual patterns. In *Proc. KDD 2007*, pp. 864–873.