

# Visualizing Sequential Patterns for Text Mining

Pak Chung Wong, Wendy Cowley, Harlan Foote, Elizabeth Jurrus, Jim Thomas

Pacific Northwest National Laboratory

{pak.wong | wendy.cowley | harlan.foote | elizabeth.jurrus | jim.thomas}@pnl.gov

## Abstract

*A sequential pattern in data mining is a finite series of elements such as  $A \rightarrow B \rightarrow C \rightarrow D$  where  $A$ ,  $B$ ,  $C$ , and  $D$  are elements of the same domain. The mining of sequential patterns is designed to find patterns of discrete events that frequently happen in the same arrangement along a timeline. Like association and clustering, the mining of sequential patterns is among the most popular knowledge discovery techniques that apply statistical measures to extract useful information from large datasets. As our computers become more powerful, we are able to mine bigger datasets and obtain hundreds of thousands of sequential patterns in full detail. With this vast amount of data, we argue that neither data mining nor visualization by itself can manage the information and reflect the knowledge effectively. Subsequently, we apply visualization to augment data mining in a study of sequential patterns in large text corpora. The result shows that we can learn more and more quickly in an integrated visual data-mining environment.*

## 1 Introduction

The task of sequential patterns in knowledge discovery and data mining is to identify the item that frequently precedes another item. Generally a sequential pattern can be described as a finite series of elements such as  $A \rightarrow B \rightarrow C \rightarrow D$  where  $A$ ,  $B$ ,  $C$ , and  $D$  are elements of the same domain. Each sequential pattern in data mining comes with a minimum *support* value, which indicates the percentage of total records that contain the pattern. An arbitrary example of a sequential pattern is *90% of the die-hard fans who saw the movie Titanic went on to buy the movie sound track CD, followed by the videotape when it was released.*

The primary goal of sequential pattern discovery is to assess the evolution of events against a measured timeline and detect changes that might occur coincidentally. This information has been used to detect

medical fraud in insurance claims, evaluate drug performances in pharmaceutical industry, and determine risk factors in military operations. Although the definition of sequential patterns varies slightly among different disciplines and applications, they all share a similar goal that can be traced back to the groundbreaking work by Agrawal and Srikant [1].

This paper presents a powerful visual data-mining [13] system that contains a customized sequential pattern discovery engine and an interactive visualization tool. The system was developed to support our ongoing text mining and visualization research [6][8][10][12][14] on large unstructured document corpora. The objective is to discover inter-transaction patterns such that the presence of a topic is followed by another topic. We show that we can learn more and more quickly by augmenting the data-mining process with visualization in an integrated visual data-mining [13] environment. We demonstrate the results using a time-series corpus that spans a 7-year period.

## 2 Related Work

In an earlier paper, we presented an original technique to visualize topic association rules in text mining [14]. An association rule is an implication of the form  $X \rightarrow Y$  where  $X$  is a set of antecedent items and  $Y$  is the consequent item. Given a domain set of elements, for example  $\{A, B, C, D\}$ ,  $A \rightarrow B \rightarrow C \rightarrow D$  is a sequential pattern while  $A + B + C \rightarrow D$  is an association rule. A sequential pattern is the study of “ordering” or “arrangement” of elements, whereas an association rule is the study of “togetherness” of elements. Some of our other text mining and visualization work includes thematic analysis of long articles using orthogonal wavelets [8] as well as topic clustering of corpora based on multidimensional scaling and other statistical measures [12].

Conventional sequential pattern mining work on relational databases can be found in the classic work by Srikant and Agrawal [1][9]. Unfortunately, visualization is not emphasized in their work.

MineSet by Silicon Graphics has perhaps the best balance between powerful knowledge discovery functions and an impressive 3D interactive visualization [2][3] for more than a dozen mining tasks. However, sequential pattern discovery is not supported.

Westphal and Blaxton [11] give an excellent introduction to state-of-the-art visualization and data mining techniques available in both commercial and public domain packages.

### 3 Data Description and Pre-Processing

The experimental results and graphics presented in Section 4 are generated using a corpus obtained from open sources. The medium-sized (~1MB) corpus is stored as an ASCII file with about 1,170 articles collected from 1991 to 1997. This corpus has a strong theme associated with nuclear smuggling news throughout the 90s. The second dataset is a news corpus harvested from the TREC5 distribution disks. It covers the entire year of 1990 including events such as the Iraq-Kuwait conflict that led to Desert Storm.

#### 3.1 Topic Extraction

The first step in processing the corpus is to identify an interesting set of content-bearing words from the articles. Words separated by white spaces in a corpus are evaluated within the context of the corpus to assess whether a word is interesting enough to be a topic. Bookstein's [4] ideas regarding identification of content-bearing words are used to assess the relative contribution of a word to the content of the corpus. The co-occurrence or lack of co-occurrence of these interesting words in documents is used to evaluate the strengths of the words. Stemming is used to remove suffixes so that similar words are represented by the root word. Commonly appearing words that do not directly contribute to the content – such as prepositions, pronouns, adjectives, and gerunds – are ignored. The result is a set of content-bearing words that represent the entire major topics (or themes) of the corpus.

#### 3.2 Multiresolution Binning

Because our primary goal is to study sequential patterns of the daily events recorded in a corpus, every topic word that appears in articles with the same time stamp is binned together in a topic subset for the mining task. Similarly, we can bin the topics by weeks, months, or years to show different resolutions of sequential patterns. The idea is to capture the topic patterns of the news stories that span different time intervals such as days, months, or years.

### 4 Discovery of Sequential Patterns

Although different applications have varying definitions of “sequential patterns,” they all share the same basic intention to find the repeating patterns of discrete events along a timeline. In the study of market basket databases, the goal is to track individual buyers' purchasing patterns [1]. The intent of our investigation, however, is to find sequential patterns of major topics extracted from a corpus. We show that either visualization or data mining alone can carry out the discovery task with a certain degree of success up to a certain point.

#### 4.1 Discovery by Visualization

We present a simple visualization example to highlight the strengths and weaknesses of using visualization to discover sequential patterns. Figure 1 is a plot of “topic combinations” versus time that we discussed in [6]. If we replace “topic combinations” with individual topics in Figure 1, we can spot the sequential patterns of individual topics by comparing the icons of individual columns along the timeline.

##### 4.1.1 Strength of Visualization

The strength of this visualization approach is that we can quickly obtain an overall structural view of topic patterns and their distribution. We can see not only the frequency of the patterns but also the occurrence dates of individual events.

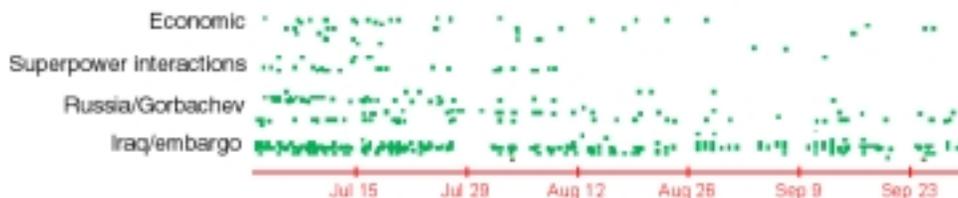


Figure 1: A plot of topic combinations of interest versus time from July to September 1990.

### 4.1.2 Weakness of Visualization

A major weakness of this visualization approach is the precision of the pattern. We do not know the exact connections of a pattern. A second weakness is the lack of statistical support of individual patterns. Because we can only display a portion of the corpus at a time, it is impossible to predict the distribution and concentration of the patterns when they are not displayed on screen. These problems, however, can be solved by including a data mining step ahead of the visualization.

## 4.2 Discovery by Data Mining

The discovery of sequential patterns plays an important role in today's data mining industry [11]. Almost every application, however, has its own specific definition and requirements. For example, we treat a break in the input data as a break in the patterns. A possible alternative is to simply accept a wider time gap. Instead of adding more parameters to our system, we choose to provide a multiresolution approach with different binning sizes to provide a coarser pattern that might otherwise be broken up by a single daybreak.

Our customized text-mining engine builds the pattern structure using an  $n$ -ary tree with  $n$  branches, also known as a *trie* in the study of computer algorithms.

Each node of the tree represents an element, i.e., a topic in our case, of the pattern sequences. The pattern is a valid pattern if its support value is larger than a pre-defined threshold value. The support value is calculated as the number of occurrences of the pattern in the dataset. Elements of a pattern are in sequential records in the input. A break in the input data represents a break in the rules.

Figure 2 shows a basic example of mining sequential patterns of a corpus with a topic domain of {A, B, C} within a 6-day period. During the first phase, we build all the patterns with 2 topics. For example, the pattern  $A \rightarrow B$  happens in Day 2 to Day 3 and Day 5 to Day 6; the support is  $2/6 = 0.33$ .

We then calculate the frequency of each pattern and prune the leaves of the trees with support values less than the threshold, i.e., 10% or 0.1 in this case. Because an  $n$ -topic sequential pattern must have the same or higher support than any patterns with the same  $n$  leading topics, we can prune the leaves of a trie as soon as they fail the threshold test. For example, the 2-topic pattern  $A \rightarrow C$  appears at least as often as any patterns with the same two leading topics such as  $A \rightarrow C \rightarrow B$ . If the support of  $A \rightarrow C$  is below the threshold, the support of  $A \rightarrow C \rightarrow B$  will only be the same or lower.

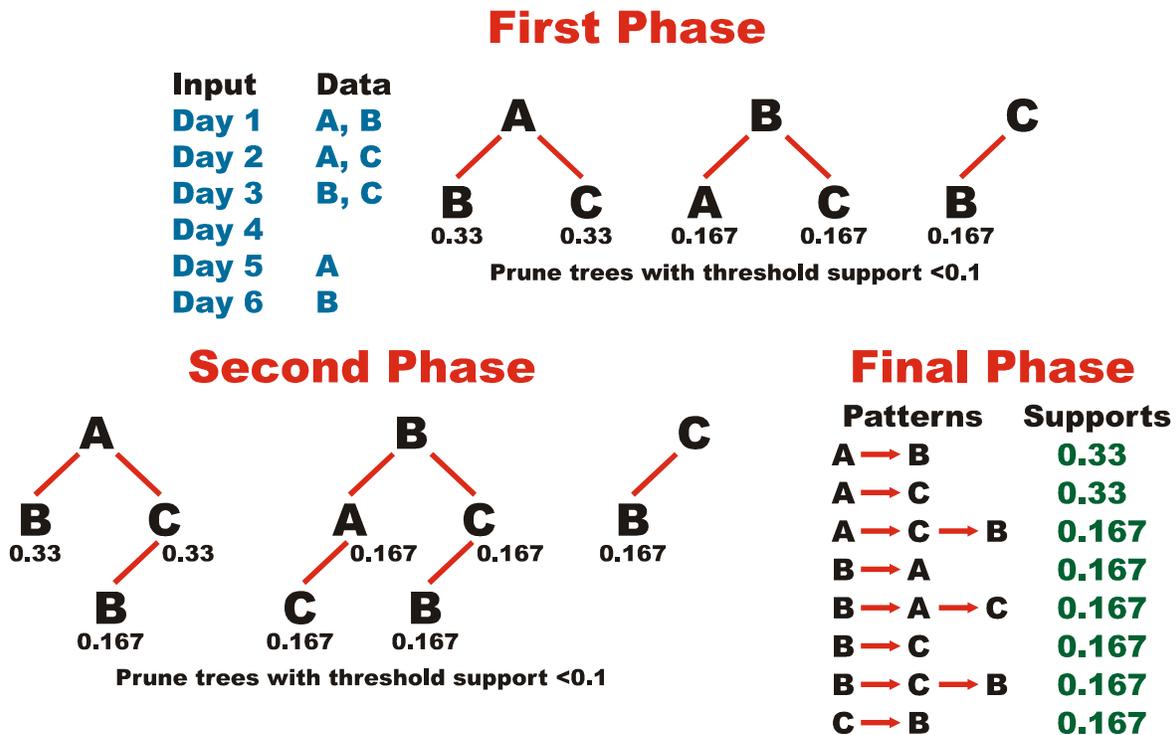


Figure 2: An example to illustrate the mining of sequential patterns from a small time-series dataset. The input data are blue. The output patterns and their supports are green. The mining task completes after two phases of matching and pruning.

In the second phase, we match the qualified 2-topic patterns with the input data and generate the new 3-topic patterns followed by another round of threshold pruning. This process is repeated with a new topic appended to the previously qualified patterns every time, until the input data are exhausted. In the example shown in Figure 2, the process stops after two phases.

The patterns are listed in the final phase with their support values. The pattern  $A \rightarrow C$  happens in Day 1 to 2 and Day 2 to 3; the support is  $2/6 = 0.33$ . The pattern  $A \rightarrow C \rightarrow B$  only happens once; the support is  $1/6 = 0.167$ .

#### 4.2.1 Performance and Statistics

We study the performance of our data mining engine using the nuclear smuggling corpus described in Section 3. The topics of the corpus are binned into topic subsets. There are 766 daily bins, 76 monthly bins, and 7 yearly bins. The last (yearly) case is merely for the sake of completeness. We really cannot learn anything from a domain size of 7. The runtime parameters and statistics of the mining process are listed in Table 1. With proper pruning and realistic support thresholds, we encounter no problem processing the corpus on a standard SUN ULTRA 10.

**Table 1: Runtime Statistics and Results**

	Number of Bins	Support Threshold	Number of Patterns
Daily	766	1%	91
Monthly	76	10%	444,617
Monthly	76	25%	509
Yearly	7	45%	148,667

#### 4.2.2 Strength of Data Mining

The strength of the data mining approach is that it provides accurate support information for all sequential patterns as well as their sub-patterns. We can study not only the strongest patterns but also the weaker ones and those in between. This can compensate for the lack of support information in the previous visualization approach.

#### 4.2.3 Weakness of Data Mining

A major weakness of the data mining approach is that we lose the temporary distribution of individual sequences. We only know there is a strong pattern, but we do not know whether it happens frequently within a short period of time or if it spreads evenly along the timeline. This locality information is particularly important if we analyze a large corpus that spans multiple years. A simple visualization that is similar to the one

in Figure 1 or the one presented in Section 5 can solve this problem. We simply plot the patterns along the timeline whenever it happens.

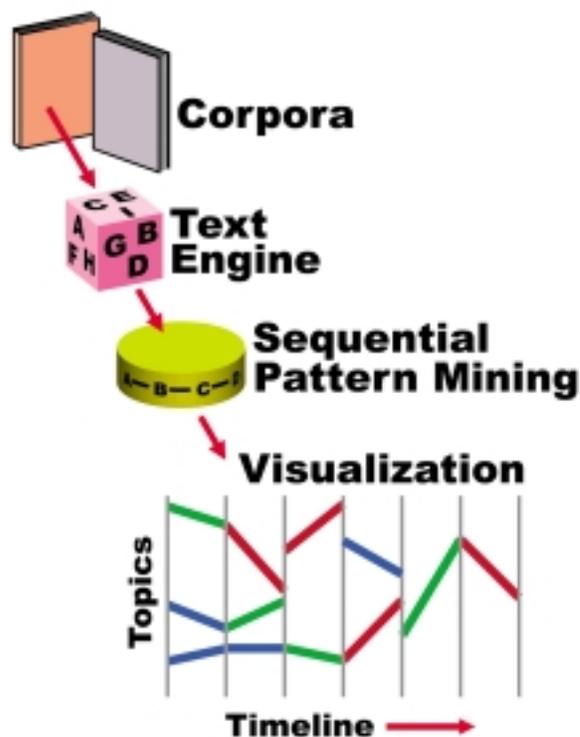
A second weakness is the size of the text output. In our experiments, we can easily generate hundreds of thousands of patterns from a medium-size corpus with a few thousand articles. A picture may not be worth a thousand words, but humans do learn faster from a visual display than a text document filled with sequences.

## 5 Visual Data Mining

Our discussion in Section 4 suggests that the strengths of visualization and data mining can be used to compensate for each other's weaknesses in the discovery of sequential patterns. We present a visual data mining system that uses visualization to augment the data mining process and takes advantage of the strengths of both worlds.

### 5.1 System Overview

Figure 3 shows a high-level overview of our visual data mining system in the discovery of sequential pat-



**Figure 3: An overview of our visual data mining system in the discovery of sequential patterns of a corpus.**

terns. A corpus of narrative text is fed into a text engine for topic extraction. Selected topics in each article are binned together for the mining engine described in Section 4 to generate the sequential patterns with their support information. Finally, the patterns are sent to the visualization system for further analysis.

## 5.2 Visualization of Sequential Patterns

We present a novel technique to visualize pattern sequences as well as their strengths along a time line. The Java-based system can be executed on multiple platforms including PC, Mac, and UNIX. The basic visualization design is sketched at the bottom of Figure 3. Topics are listed on the left-hand side.

Each vertical axis represents the topics of one time step. Each connecting line represents a sequential pattern. Colors are used to show the support of the patterns as well as sub-patterns. The appearance of our visualization is somewhat similar to parallel coordinates [5][7]. The horizontal axis (time) has a natural ordering, but the ordering of the elements on the vertical axis (topics) is essentially arbitrary. This is different from the situation in parallel coordinates where the ordering of the horizontal parallel axes is arbitrary, and the ordering of the elements on the vertical axes is given.

Figure 4 shows a snapshot of the prototype system. The vertical lines are turned off to simplify the display. The beginning of a pattern is marked by a black dot.

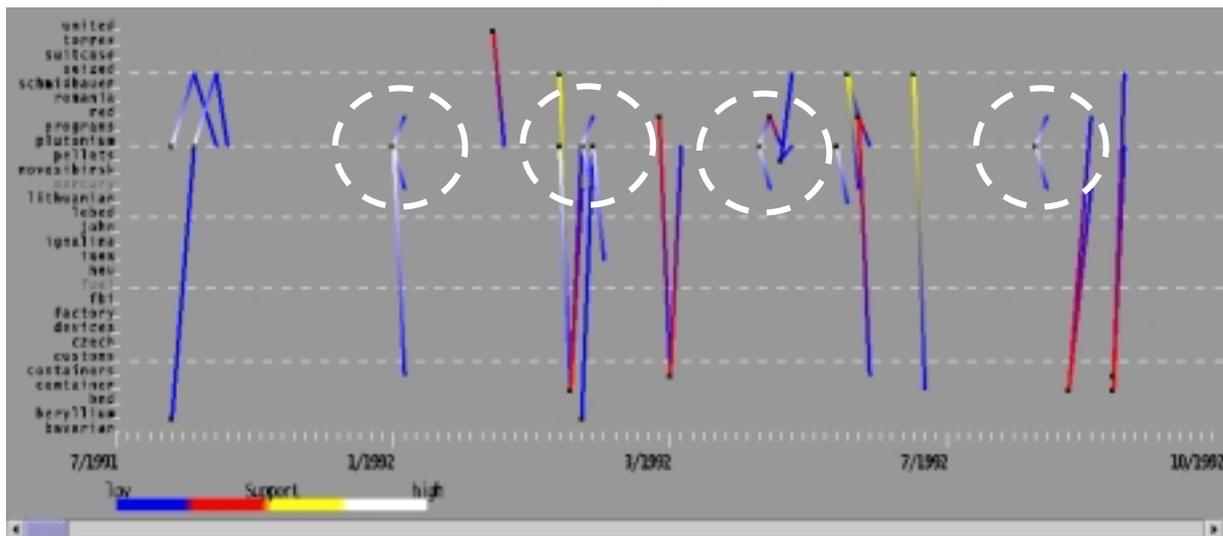


Figure 4: A snapshot of the visualization system. The topics are listed on the left-hand side. Colored lines are used to connect the topics to form a sequential pattern. The beginning of a pattern is marked by a black dot. Four basic colors – shown at the bottom – are used to indicate the support of a pattern and its sub-patterns. Shading is turned on to blend the colors between two sub-pattern support values. The four white dashed circles mark the four appearances of the same two patterns within an 8-month period from Jan 92 to Aug 92.

Four basic colors – blue, red, yellow, and white – are used to show different levels of support values that are normalized for visualization. Shading is turned on to blend the colors between two consecutive topics of a pattern. The four white dashed circles highlight the four appearances of the same two pattern sequences “plutonium → red” and “plutonium → mercury”. These patterns came from the articles recorded right after the collapse of the former Soviet Union in 1992. Western European officials frequently seized plutonium and a substance known as “red mercury” said to be smuggled from the former Soviet Union. The mysterious “red mercury”, which was reportedly peddled throughout Europe and the Middle East by Russian businessmen who made fortunes, turned out to be a hoax.

## 5.3 Multiresolution Visualization and Query

Figure 5 depicts another visualization using the same corpus but with different binning resolution. This time we bin the topics of the articles monthly instead of daily. As the figure shows, this can drastically increase the number of sequential patterns in our visualization because we now have fewer bins (76) but each bin has a lot more topics accumulated. In Table 1, we show that we can get as many as 444K patterns using 10% support threshold. A quick way to clear up this visualization is to increase the threshold values. The number of patterns drops from 444K to less than 600 after we up the threshold from 10% to 25%. The visualization in

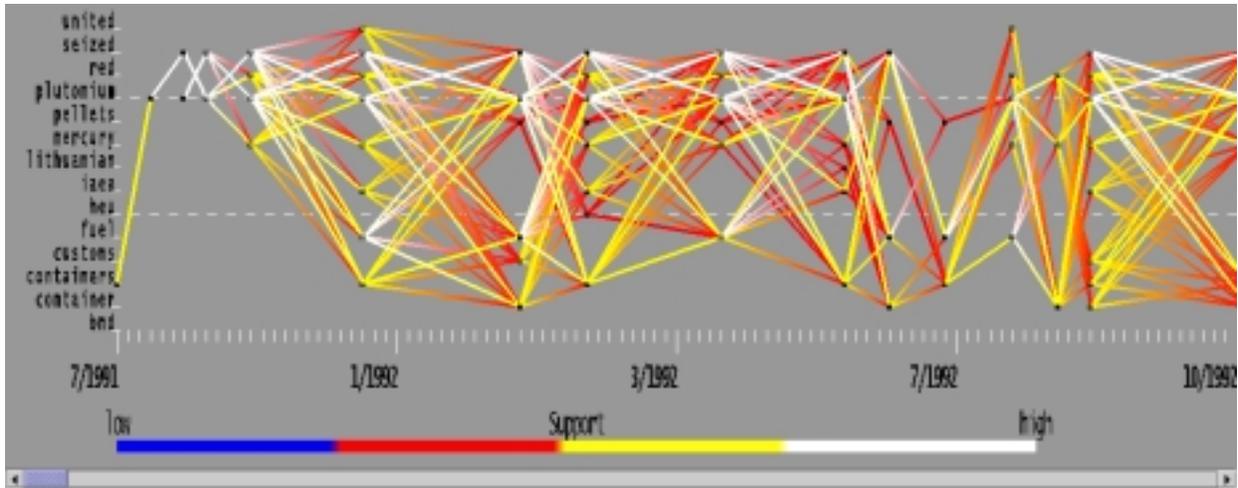


Figure 5: A visualization snapshot of sequential patterns using the monthly binning topic sets.

Figure 5 has a minimum threshold value of 25%. Another practical option provided by our prototype is the topic query capability to select patterns with particular topics. A pop-up topic selection window, as shown in Figure 6, allows users to pick patterns associated with an individual topic or a set of topics.

## 6 Discussion

In data mining studies, a lot of research activities involve the integration of data mining and databases. The idea is to provide selection capabilities to extract knowledge from huge databases using SQL or other query languages. In information visualization, numerous multiresolution efforts – such as the context/focus approach or the overall-structure/local-details approach – help humans see the whole and select the parts. Both communities are trying to find an effective way to extract knowledge from massive amounts of data by eliminating unwanted and usually noisy information.

Without the statistics measures to reflect the support of a pattern, the visualization would be meaningless. We simply don't know if a sequence is strong enough to be a "pattern" even though we can spot it. With the graphical encoding of visualization, the pattern mining approach would be enhanced with both spatial and

- bavarian
- beryllium
- bnd
- container
- containers
- customs
- czech
- devices
- factory
- fbi
- fuel
- heu
- iaea
- ignalina
- john
- lebed
- lithuanian
- mercury
- novosibirsk
- pellets
- plutonium
- programs
- red
- romania
- schmidbauer
- seized
- suitcase
- torres
- united

Figure 6: A topic selection window.

temporal information that in turn help humans to interpret the mining results. We show that, in this particular case, it is better to combine the statistical efforts of data mining with the human participation of visualization to achieve the ultimate goal of knowledge discovery.

Our visual data mining system described in this paper is neither perfect nor complete. It would be helpful to have secondary memory support for our data-mining engine so that we can handle even larger datasets. It would also help to increase the display volume of our visualization system; a pixel-based visualization design may help to accomplish this goal. Perhaps using visualization to enhance the topic-matching and tree-pruning algorithm of a data mining process would be an even stronger visual data-mining paradigm [13].

## 7 Conclusions

This paper presents data mining and visualization techniques for discovery of sequential patterns from large datasets. We conclude that the strengths of the two approaches can compensate for each other's weaknesses. We then introduce a powerful visual data-mining environment that contains a data-mining engine to discover the patterns and their support values and a visualization front-end to show the distribution and locality of the patterns. Our result shows that we can learn more and more quickly in such an integrated visual data-mining environment.

Our long-term goal is to integrate many of our tools and techniques such as association rules [14] into a single visual data mining environment that provides user-friendly navigation, in-depth association and implication analysis, time sequence analysis, hypothesis explanation, and document summarization.

## Acknowledgments

The Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute under contract DE-AC06-76RLO 1830. This research has been supported by a Laboratory Directed Research and Development grant funded by the U.S. Department of Energy for the Pacific Northwest National Laboratory. We wish to thank Dan Adams, George Chin, Kris Cook, Sharon Eaton, Beth Hetzler, Wanda Mar, Dennis McQuerry, Ted Tanasse, and Paul Whitney who provided assistance of many forms throughout this research.

## References

- [1] Rakesh Agrawal and Ramakrishnan Srikant. Mining Sequential Patterns. In *Proceedings of the International Conference on Data Engineering (ICDE)*, Taipei, Taiwan, March 1995.
- [2] Barry G. Becker. Volume Rendering for Relational Data. In John Dill and Nahum Gershon, editors, *Proceedings Information Visualization '97*, pages 87 – 90, Los Alamitos, CA, Oct 20 – 21, 1997. IEEE CS Press.
- [3] Barry G. Becker. Visualizing Decision Table Classifiers. In Graham Wills and John Dills, editors, *Proceedings of Information Visualization '98*, pages 102-105, Los Alamitos, CA, Oct 19 – 20, 1998. IEEE CS Press.
- [4] A. Bookstein, S.T. Klein, and T. Raita. Clumping Properties or Content-Bearing Words. *Journal of the American Society for Information Science*, 49(2):102 – 114, 1998.
- [5] Ying-Huey Fua, Matthew O. Ward, and Elke A. Rundensteiner. Hierarchical Parallel Coordinates for Exploration of Large Datasets. In David Ebert, Markus Gross, and Bernd Hamann, editors, *Proceedings IEEE Visualization '99*, pages 43 – 50, New York, NY, Oct 24 – Oct 29, 1999. ACM Press.
- [6] Beth Hetzler, Paul Whitney, Lou Martucci, and Jim Thomas. Multi-faceted Insight through Interoperable Visual Information Analysis Paradigms. In Graham Wills and John Dill, editors, *Proceedings Information Visualization '98*, pages 137 – 144, Los Alamitos, CA, Oct 19-20, 1998. IEEE CS Press.
- [7] Alfred Inselberg and Bernard Dimsdale. Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry. In Arie Kaufman, editor, *Proceedings IEEE Visualization '90*, pages 361 – 375, Los Alamitos, CA, Oct 1990. IEEE Computer Society Press.
- [8] Nancy E. Miller, Pak Chung Wong, Mary Brewster, and Harlan Foote. TOPIC ISLANDS™ - A Wavelet-Based Text Visualization System. In David Ebert, Hans Hagan, and Holly Rushmeier, editors, *Proceedings IEEE Visualization '98*, pages 189 – 196, New York, NY, Oct 18 – 23, 1998. ACM Press.
- [9] Ramakrishnan Srikant and Rakesh Agrawal. Mining Sequential Patterns: Generalizations and Performance Improvements. In *Proceedings the Fifth International Conference on Extending Database Technology (EDBT)*, Avignon, France, March 1996.
- [10] Jim Thomas, Kris Cook, Vern Crow, Beth Hetzler, Richard May, Dennis McQuerry, Renie McVeety, Nancy Miller, Grant Nakamura, Lucy Nowell, Paul Whitney, and Pak Chung Wong. Human Computer Interaction with Global Information Spaces – Beyond Data Mining. In *Proceedings British Computer Society Conference*, Bradford, UK, April 1999. Springer Verlag.
- [11] Christopher Westphal and Teresa Blaxton. *Data mining solutions – Methods and Tools for Solving Real-Word Problems*, New York, 1998. John Wiley and Sons.
- [12] James A. Wise, James J. Thomas, Kelly Pennock, David Lantrip, Marc Pottier, Anne Schur, and Vern Crow. Visualizing the Non-visual: Spatial Analysis and Interaction with Information from Text Documents. In Nahum Gershon and Steve Eick, editors, *Proceedings IEEE Information Visualization '95*, pages 51 – 58, Los Alamitos, CA, Oct 20 – 21, 1995. IEEE CS Press.
- [13] Pak Chung Wong. Visual Data Mining – Guest Editor's Introduction. *IEEE Computer Graphics and Applications*, Vol 19, No 5, Los Alamitos, CA, 1999. IEEE CS Press.
- [14] Pak Chung Wong, Paul Whitney, and Jim Thomas. Visualizing Association Rules for Text Mining. In Graham Wills and Daniel Keim, editors, *Proceedings of IEEE Information Visualization '99*, Los Alamitos, CA, 1999. IEEE CS Press.