Visualizing Association Rules for Text Mining

Pak Chung Wong, Paul Whitney, Jim Thomas
Pacific Northwest National Laboratory†

ABSTRACT

An association rule in data mining is an implication of the form $\boldsymbol{X} \rightarrow Y$ where $\boldsymbol{X}$ is a set of antecedent items and Y is the consequent item. For years researchers have developed many tools to visualize association rules. However, few of these tools can handle more than dozens of rules, and none of them can effectively manage rules with multiple antecedents. Thus, it is extremely difficult to visualize and understand the association information of a large data set even when all the rules are available. This paper presents a novel visualization technique to tackle many of these problems. We apply the technology to a text mining study on large corpora. The results indicate that our design can easily handle hundreds of multiple antecedent association rules in a three-dimensional display with minimum human interaction, low occlusion percentage, and no screen swapping.

1 INTRODUCTION

Association is a powerful data analysis technique that appears frequently in data mining literature. An association rule is an implication of the form $\boldsymbol{X} \rightarrow Y$ where $\boldsymbol{X}$ is a set of antecedent items and Y is the consequent item. An example association rule of a supermarket database is *80% of the people who buy diapers and baby powder also buy baby oil*. The analysis of association rules is used in a variety of ways, including merchandise stocking, insurance fraud investigation, and climate prediction. For years scientists and engineers have developed many visualization techniques to support the analyses of association rules. Commercial data mining systems such as SGI's MineSet [10] and IBM's QUEST [8] provide tools to visualize the associations of business databases. Many of the visualizations, however, have come up short in dealing with numerous rules or rules with multiple antecedents. This limitation presents serious challenges for analysts who need to understand the association information of large databases.

This paper presents a novel association-rule visualization system designed to tackle many of these problems. The system was developed to support our ongoing text mining and visualization research [6][7][11][13][14][15] on large unstructured document corpora. The focus is to study the relationships and implications among *topics*, or descriptive concepts, that are used to characterize a corpus. The goal is to discover important association rules within a corpus such that the presence of a set of topics in an article implies the presence of another topic. For example, one might learn in headline news that whenever the words "Greenspan" and "inflation" occur, it is highly probably that the stock market is also mentioned. We demonstrate the results using a news corpus with more than 3,000 articles collected from open sources. We show that it is critical to have an effective visualization tool to support the analysis of topic associations on large corpora.

2 RELATED WORK

A visualization of association rules is a depiction of one-to-one or many-to-one mapping of information items. Prior work on visualization of association rules can be found in commercial data mining software such as MineSet [10] and QUEST [8][9]. The matrix-based visualization designs that position items on separate axes are among the most popular approaches to visualize binary relationships. Hetzler et al. [6] animate a directed graph to visualize the associations and disassociations of information items. Becker [1][2] describes a series of elegant visualization techniques designed to support data mining of business databases. Westphal et al. [16] give an excellent introduction of visualization techniques provided by current data mining tools.

† P.O. Box 999, Richland, WA 99352, USA.
pak.wong@pnl.gov, paul.whitney@pnl.gov, jim.thomas@pnl.gov
Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute under contract DE-AC06-76RLO 1830.

3 ASSOCIATION RULE DEFINITION

The definition of an association rule varies with disciplines and implementations. Our definition of association rules is similar to that found in QUEST [9]. The basic approach is to mine *qualitative* [12] rules that describe associations between sets of items. This is different from the mining of *quantitative* rules in scientific simulation and modeling. Because our application (text analysis) requires no domain knowledge for mining, we adopt the qualitative definition in our implementation.

Given a set of items, $\boldsymbol{S} = \{i_1, i_{2,...}, i_j, \ldots, i_n\}$ where $n \geq 2$. An association rule is an implication of the form $\boldsymbol{X} \rightarrow i_j$ where $\boldsymbol{X} \subset \boldsymbol{S}$, and $i_j \in \boldsymbol{S}$ such that $i_j \notin \boldsymbol{X}$. The set of items $\boldsymbol{X}$ is the antecedent, while the item i_j is the consequent of the association rule. The size of $\boldsymbol{X}$ is between 1 to (n-1) items. The *support* of the rule $\boldsymbol{X} \rightarrow i_j$ is defined as the percentage of items in $\boldsymbol{S}$ that satisfies the union of items in $\boldsymbol{X}$ and i_j. The *confidence* of the rule is the percentage of articles that satisfies $\boldsymbol{X}$ and also satisfies i_j. MineSet [10] has similar definitions but uses the terms *predictability* and *prevalence* (instead of confidence and support) to describe the strengths of association rules. These are also naturally interpreted as estimates of *conditional* and *joint* probability.

4 ASSOCIATION RULE VISUALIZATION

At least five parameters are involved in a visualization of association rules: sets of antecedent items, consequent items, associations between antecedent and consequent, rules' support, and confidence. Our goal is to visualize a large number of association rules and their metadata in a three-dimensional (3D) display with minimum human interaction, minimum occlusion, and no screen swapping. There is no maximum limit on the number of antecedent items allowed in an association.

4.1 Current Technology

The two prevailing approaches used today to visualize association rules are the *two-dimensional matrix* and *directed graph*. This section describes the general design of each approach and discusses their strengths and weakness.

4.1.1 Two-Dimensional Matrix

The basic design of a two-dimensional (2D) association matrix positions the antecedent and consequent items on separate axes of a square matrix. Customized icons are drawn on certain matrix tiles that connect the antecedent and the consequent items of the corresponding association rules. Different icons can be used to depict different metadata such as the support and confidence values of the rules. Figure 1 depicts an association rule (B→C). Both the height and the color of the column icon can

Figure 1: B→C.

be used to present metadata values. In Quest [8], the values of support and confidence are mapped to 3D columns that are built separately on and beneath the matrix tiles. Other icons such as disk and bar are also used to visualize metadata in the Rule Visualizer of MineSet [10].

A 2D matrix is arguably the most effective technique to show *one-to-one* binary relationship [17]. It has a long history of analyzing a wide variety of data in different disciplines. The strengths of a 2D matrix, however, break down when we need to visualize *many-to-one* relationships such as association rules with multiple antecedent items. For example, in Figure 2 it is almost impossible to tell whether there is only one association rule (A+B→C) or two (A→C and B→C). The lack of a practical way to identify the togetherness of individual antecedent items makes a 2D matrix a weaker candidate to visualize rules with multiple antecedent items.

Figure 2: (A+B→C) or (A→C and B→C).

MineSet addresses the problem by grouping all the antecedent items of an association rule as one unit and plotting it against its consequent, i.e., an antecedent-to-consequent plot. For example, a dedicated item group (A+B) is created in Figure 3 to describe the association rule (A+B→C). The strategy works fine for smaller antecedent sets (e.g., less than 3 items). In our text mining studies, we encounter association rules with as many as 12 items in the antecedent. The replication of items in the antecedent groups creates a much larger antecedent-to-consequent plot when compared with the corresponding item-to-item plot. The loss of item identity within an antecedent group also defeats the purpose of visualizing the associations with a matrix. For example, the row (or column) of the matrix connected to an item can no longer be used to search for all the rules involving that item. Another problem in a 2D-matrix display is object occlusion, especially when multiple icons are used to depict different metadata values on the matrix tiles. The occlusion problem is obvious in Figure 4.

Figure 3: Identities of A and B are lost in (A+B).

Figure 4: Object occlusions.

4.1.2 Directed Graph

A directed graph is another prevailing technique to depict item associations. The nodes of a directed graph represent the items, and the edges represent the associations. Figure 5 shows three association rules (A→C, B→C, A+B→C). This technique works well when only a few items (nodes) and associations (edges) are involved. An association graph can quickly turn into a tangled display with as few as a dozen rules. Hetzler et al. [6] address the problem by animating the edges to show the associations of certain items with 3D rainbow arcs. The animation technique requires significant human interaction to turn on and off the item nodes. It is not an easy task to show multiple metadata values, including support and confidence, alongside the association rules.

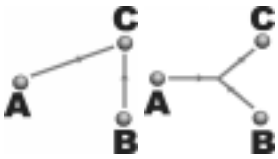

Figure 5: Left: (A→C and B→C). Right: A+B→C.

4.2 A Novel Visualization Technique

We present a novel technique to visualize *many-to-one* association rules. Instead of using the tiles of a 2D matrix to show the *item-to-item* association rules, we use the matrix to depict the *rule-to-item* relationship. In Figure 6 and CP1, the rows of the matrix floor represent the items (or *topics* in the context of text mining), and the columns represent the item associations. The blue and red blocks of each column (rule) represent the antecedent and the consequent of the rule. The identities of the items are shown along the right side of the matrix. The confidence and support levels of the rules are given by the corresponding bar charts in different scales at the far end of the matrix. The system supports basic query commands through the use of pop-up menus to restrict key items to be included in the visualization. The display has a mouse-controlled zooming capability to support context/focus analysis.

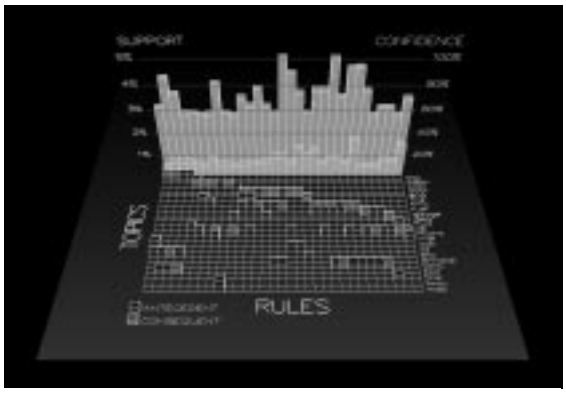

Figure 6: A visualization of item associations. See CP1.

The rule-to-item visualization approach has many advantages over all the other matrix-based predecessors:

- There is virtually no upper limit on the number of items in an antecedent.
- We can analyze the distributions of the associations (horizontal) and the items within (vertical) simultaneously.
- Unlike Figure 3, the identity of individual items within an antecedent group is clearly shown.
- No new antecedent groups are created because of the multiple antecedent items in association rules.
- Because all the metadata are plotted at the far end and the heights of the columns are scaled, few occlusions occur.
- No screen swapping, animation, or human interaction (other than basic mouse zooming) is required to analyze the rules.

From our experiments, this technique works well with up to several hundred association rules and dozens of antecedent topics on a 17-inch monitor. This is a significant improvement over the traditional 2D item-to-item association visualization.

5 TEXT MINING - AN APPLICATION

This section briefly describes our design philosophy and implementation issues of a text association mining system. (See [18] for more information about the overall system.) The design is based on ideas from information retrieval and syntactic analysis. The underlying text engines, like others [4][5], construct a mathematical signature of a block of text. If similar text corresponds with similar signatures, then some useful text analyses can be performed on the signatures using standard mathematical and data analytic techniques.

5.1 Demonstration Corpus

The experimental results and graphics presented in this paper are generated using a news article corpus obtained from open sources. The medium-sized (~9MB) news corpus is stored as an ASCII file with more than 3,000 articles collected during April 20-26, 1995. This corpus has a strong theme associated with the bombing of the U.S. Federal Building in Oklahoma. Other major news stories during the week include the Simpson trial, the Unabomber, the Bosnian crisis, and the France election.

5.2 System Overview

Figure 7 shows a high-level system overview of the topic association mining system. A corpus of narrative text is fed into a text engine for topic extractions. The mining engine then reads the topics from the text engine and generates topic association rules. Finally, the resultant association rules are sent to the visualization system for further analysis.

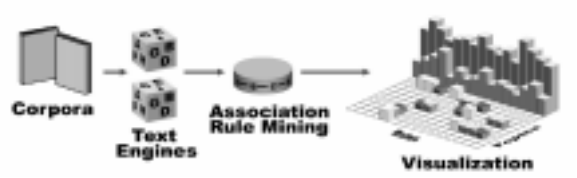

Figure 7: Overview of the topic association system.

5.3 Text Engine

We developed two text engines to generate conceptual topics from a large corpus. The first one is *word*-based and results in a list of content-bearing words for the corpus. The second is *concept*-based and results in concepts (represented as word groups) based on the corpus. The engines are similar in that the topics and concepts are initially evaluated using the entire corpus. Because of limited space, we only present the word-based engine here. A description of the concept-based text engine is available in [18].

The word-based text engine selects an interesting subset of words. Words separated by white spaces in a corpus are evaluated *within the context of that corpus* to assess whether a word is "interesting" enough to be a topic. Bookstein's [3] 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.

Table 1 shows the top 20 topic words generated from this text engine using the news corpus. The content-bearing topic words (all nouns in this example) included in the table represent many familiar headline news stories of the week.

Table 1: Top 20 topic words and their corresponding news stories.

Topic Word	News Story
billion	Congress and budget
Neufeld	Simpson trial
Korea	North/South Korea nuclear talks
Chirac	France election
tribunal	Bosnian crisis
Bosnian	Bosnian crisis
Unabomber	Unabomber
CIA	Unabomber
Mazzola	Simpson trial
treaty	North/South Korea nuclear talks
Jospin	France election
Serb	Bosnian crisis
McVeigh	Oklahoma bombing
Simpson	Simpson trial
Nichols	Oklahoma bombing
Nuclear	North/South Korea nuclear talks
Sarajevo	Bosnian crisis
Ito	Simpson trial
Koernke	Oklahoma bombing
refugee	Bosnian crisis

5.4 Topic Association Mining

The topic words selected by the text engine are fed into the mining engine to compute the association rules according to the requested confidence and support values. Table 2 shows a sample of topic association rules with confidence ≥ 80% and support ≥ 1% generated from the April 1995 news corpus.

Table 2: A sample of twelve association rules (not in order).

Antecedent	Consequent	Confidence	Support
Manager & McVeigh & Michigan & motel & Nichols & sketch	Truck	91.30%	1.39%
Ito & court & jury & Mazzola & testimony	Simpson	100.00%	1.72%
France & election & socialist	Chirac	97.30%	1.19%
Blood & testimony	Fung	81.25%	1.29%
Court & judge & jury	Ito	96.49%	1.81%
Blood & vial	Mazzola	100.00%	1.19%
Ammonium & nitrate & bomb & FBI	Nichols	89.13%	1.35%
Bomb & cult & Michigan & militia	McVeigh	100.00%	1.22%
Hutu & Rwanda	Refugee	92.86%	1.29%
Bosnia & Bosnian & Crimes	Serb	100.00%	1.22%
Bosnia & Serb & Sarajevo	War	90.48%	1.25%
Cult & gas & subway	Tokyo	95.00%	1.88%

5.5 Topic Association Visualization

Figure 8 and CP 2 depict a set of topic association rules extracted from the April 1995 news corpus with support ≥ 6% and confidence ≥ 60%. The rules are shown in ascending order from the left side according to the confidence values. This visualization can be used to study different aspects of the topic associations, including the topic distribution of selected rules and the correlation between different metadata values. Figure 9 and CP3 show a different arrangement of the same set of association rules, which are now ordered by the consequent items. All the associations with the same consequent items are group together for analysis.

6 DISCUSSION

Switching from an *item-to-item* arrangement to a *rule-to-item* design of a 2D matrix lets us effectively display association rules with multiple antecedent topics. The current design works well with up to several hundred association rules on a 17-inch monitor

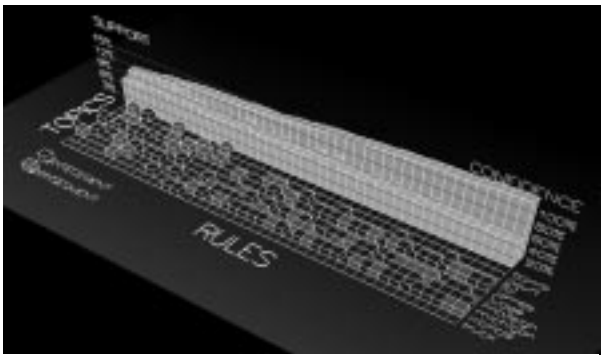

Figure 8: Rules are sorted by confidence values. See CP2.

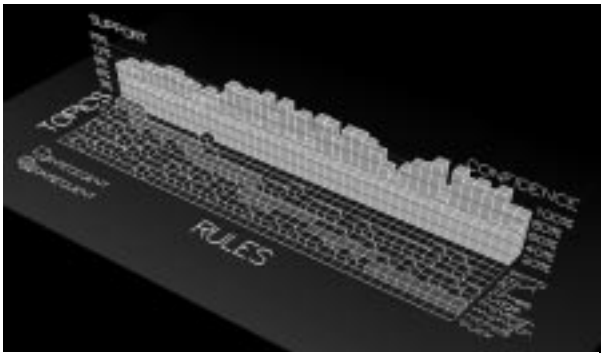

Figure 9: Rules are sorted by consequent topics. See CP3.

screen. With mouse zooming and topic selection, we encounter no problems when analyzing 10,000+ document corpora using our visualization system. For the other applications that involve thousands and thousands of association rules, a pixel-based or glyph-based visualization technique [17] might be needed to display the rules and figures using pixels or glyphs instead of blocks and charts.

7 CONCLUSIONS AND FUTURE WORK

This paper presents the prevailing techniques to visualize association rules and discusses their weakness in dealing with a large number of association rules with multiple antecedent items. We introduce a new visualization technique designed to overcome many of the shortcomings of it predecessors. We apply the new technique to a text mining system to analyze a large text corpus. The results indicate that our design can easily handle hundreds of multiple antecedent association rules in a 3D display with minimum human interactions, low occlusion percentage, and no screen swapping.

Our long-term goal is to integrate many of our tools and techniques into a single visualization environment that provides user-friendly navigation, in-depth association and implication analyses, time sequence analysis, hypothesis explanation, and document summarization.

ACKNOWLEDGEMENTS

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, Kris Cook, Wendy Cowley, Vern Crow, Don Daly, Scott Decker, Sharon Eaton, Harlan Foote, Sue Harve, Beth Hetzler, Cherry Lei, Rik Littlefield, Dennis McQuerry, Nancy Miller, Grant Nakamura, Lucy Nowell, and Renie McVeety who provided assistance of many forms throughout this research.

REFERENCES

[1] Barry G. Becker. Volume Rendering for Relational Data. In *Proceedings of Information Visualization '97*, pages 87-90, Phoenix, Arizona, Oct 20 - 21, 1997. IEEE CS Press.

[2] Barry G. Becker. Visualizing Decision Table Classifiers. In *Proceedings of Information Visualization '98*, pages 102-105, Research Triangle Park, North Carolina, Oct 19 - 20, 1998. IEEE CS Press.

[3] 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.

[4] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. In *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[5] Steven Finch. *Finding Structure in Language*, Ph.D. Dissertation, University of Edinburgh, 1993.

[6] Beth Hetzler, W. Michelle Harris, Susan Havre, and Paul Whitney. Visualizing the Full Spectrum of Document Relationships. In *Proceedings of the Fifth International Society for Knowledge Organization (ISKO) Conference*, 1998.

[7] Beth Hetzler, Paul Whitney, Lou Martucci, and Jim Thomas. Multi-faceted Insight through Interoperable Visual Information Analysis Paradigms. In *Proceedings of Information Visualization '98*, pages 137-144, Research Triangle Park, North Carolina, Oct 19-20, 1998. IEEE CS Press.

[8] www.almaden.ibm.com/cs/quest/demo/assoc/general.html

[9] www.almaden.ibm.com/cs/quest/publications.html#associations

[10] www.sgi.com/software/mineset

[11] 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 of IEEE Visualization '98*, pages 189-196, New York, NY, Oct 18-23, 1998. ACM Press.

[12] Gregory Piatetsky-Shapiro, Editor. *Knowledge Discovery in Databases*, Menlo Park, CA, 1991. AAAI Press.

[13] J. S. Risch, D. B. Rex, S. T. Dowson, T. B. Walters, R. A. May, B. D. Moon. The STARLIGHT Information Visualization System. In Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman, editors, *Readings in Information Visualization - Using Vision to Think*, 1999. Morgan Kaufmann.

[14] 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 of British Computer Society Conference*, Bradford, UK, April *1999*. Springer Verlag.

[15] 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 of IEEE Information Visualization '95*, pages 51-58, Los Alamitos, CA, Oct 20-21, 1995. IEEE CS Press.

[16] Christopher Westphal and Teresa Blaxton. Data mining solutions - Methods and Tools for Solving Real-Word Problems, New York, 1998. John Wiley and Sons.

[17] Pak Chung Wong and R. Daniel Bergeron. 30 Years of Multidimensional Multivariate Visualization. In Gregory M. Nielson, Hans Hagan, and Heinrich Muller, editors, Scientific Visualization - Overviews, Methodologies and Techniques, pages 3-33, Los Alamitos, CA, 1997. IEEE CS Press.

[18] Pak Chung Wong, Paul Whitney, and Jim Thomas. Mining of Topic Associations. Technical Report PNNL-SA-30975, Pacific Northwest National Laboratory, Richland, WA, 1999.