A Framework for Visualizing Association Mining Results

Gürdal Ertek[1] and Ayhan Demiriz[2]

[1] Sabanci University
Faculty of Engineering and Natural Sciences
Orhanli, Tuzla, 34956, Istanbul, Turkey
ertekg@sabanciuniv.edu
[2] Department of Industrial Engineering
Sakarya University
54187, Sakarya, Turkey
ademiriz@gmail.com

Abstract. Association mining is one of the most used data mining techniques due to interpretable and actionable results. In this study we propose a framework to visualize the association mining results, specifically frequent itemsets and association rules, as graphs. We demonstrate the applicability and usefulness of our approach through a Market Basket Analysis (MBA) case study where we visually explore the data mining results for a supermarket data set. In this case study we derive several interesting insights regarding the relationships among the items and suggest how they can be used as basis for decision making in retailing.

1 Introduction

Association mining is an increasingly used data mining and business tool among practitioners and business analysts [7]. Interpretable and actionable results of association mining can be considered as the major reasons for the popularity of this type of data mining tools. Association rules can be classified based on several criteria, as outlined in [11]. In this paper, we focus on single-dimensional, single-level boolean association rules, in the context of market basket analysis.

By utilizing efficient algorithms such as Apriori [2] to analyze very large transactional data -frequently from transactional sales data- will result in a large set of association rules. Commonly, finding the association rules from very large data sets is considered and emphasized as the most challenging step in association mining. Often, results are presented in a text (or table) format with some querying and sorting functionalities. The rules include "if" clauses by default. The structure of such rules is as follows: "If the customer purchases Item A, then with probability C% he/she will buy Item B."

This probability, C%, is referred to as confidence level. More formally, the confidence level can be computed as follows: $C = \frac{frequency(A \cap B)}{frequency(A)}$, where $A \cap B$ refers to the transactions that have both Item A and Item B. Confidence level is also equivalent to the conditional probability of having Item B given Item A.

Another important statistic in association mining is the support level. Support level is basically equal to the fraction of the transactions that have both Item A and Item B. Thus the support level S is computed as follows: $S = \frac{frequency(A \cap B)}{T}$ where T is equal to the total number of the transactions. Left and right hand sides of the rule are called antecedent and consequent of the rule respectively.

There exists an extensive literature where a multitude of interestingness measures for association rules are suggested [22] and efficient data mining algorithms are presented for deriving these measures. However, there is considerably less work that focuses on the interpretation of the association mining results.

Information visualization, a growing field of research in computer science [6, 8, 13], investigates ways of visually representing multi-dimensional data with the purpose of knowledge discovery. The significance and the impact of information visualization is reflected by the development and availability of highly user-friendly and successful software tools such as Miner3D [18], Spotfire [21], Advizor [1], DBMiner [11], and IBM Intelligent Miner Visualization [15].

Our motivation for this study stems from the idea that visualizing the results of association mining can help end-users significantly by enabling them to derive previously unknown insights. We provide a framework that is easy to implement (since it simply merges two existing fields of computer science) and that provides a flexible and human-centered way of discovering insights.

In spite of successful visualization tools mentioned above, the visualization of the association mining results in particular is somewhat a fertile field of study. Some of the studies done in this field are summarized in the next section. We then introduce our proposed framework in Section 3. We explain our implementation in Section 4. We report our findings from the case study in Section 5. We then conclude with future research directions in Section 6.

2 Literature Review

The visualization of association mining results has attracted attention recently, due to the invention of information visualization schemes such as parallel coordinate plots (||-coords). Here we summarize some of the studies that we believe are the most interesting.

Hofmann et al. [14] elegantly visualize association rules through Mosaic plots and Double Decker plots. While Mosaic plots allow display of association rules with two items in the antecedent, Double Decker plots enable visualization of association rules with more items in the antecedent. The interestingness measure of "differences of confidence" can be directly seen in Double Decker plots. Discovering intersection and sequential structures are also discussed.

Kopanakis and Theodoulidis [17] present several visualization schemes for displaying and exploring association rules, including bar charts, grid form models, and ||-coords. In their extensive work they also discuss a similarity based approach for the layout of multiple association rules on the screen.

Two popular approaches for visualizing association rules are summarized in [23]: The first, matrix based approach, maps items and itemsets in the antecedent

and consequent to the X and Y axes respectively. Wong et al. [23] bring an alternative to this approach by mapping rules -instead of items- to the X axis. In the second, directed graph approach, the items and rules are mapped to nodes and edges of a directed graph respectively. Our framework departs from the second approach since we map both the items *and* the rules to nodes.

3 Proposed Framework

We propose a graph-based framework to visualize and interpret the results of well-known association mining algorithms as directed graphs. In our visualizations, the items, the itemsets, and the association rules are all represented as nodes. Edges represent the links between the items and the itemsets or associations.

In visualizing frequent itemsets (Figure 1) the nodes that represent the items are shown with no color, whereas the nodes that represent the itemsets are colored reflecting the cardinality of the itemsets. For example, in our case study the lightest shade of blue denotes an itemset with two items and the darkest shade of blue denotes an itemset with four items. The sizes (the areas) of the nodes show the support levels. The directed edges symbolize which items constitute a given frequent itemset.

In visualizing association rules (Figure 4), the items are again represented by nodes without coloring, and the rules are shown by colored nodes. The node sizes (the areas) again show the support levels, but this time the node colors show the confidence levels. In our case study, the confidence levels are shown in a linearly mapped yellow-red color spectrum with the yellow representing the lowest and the red representing the highest confidence levels. The directed edges are color-coded depending on whether they are incoming or outgoing edges. Incoming edges of the rule nodes are shown in grey and outgoing are shown in black. For example, in Figure 4 the rule A01 is indeed (Item 110 $\Rightarrow$ Item 38).

The main idea in our framework is to exploit already existing graph drawing algorithms [3] and software in the information visualization literature [12] for visualization of association mining results which are generated by already existing algorithms and software in the data mining literature [11].

4 Steps in Implementing the Framework

To demonstrate our framework we have carried out a case study using a real word data set from retail industry which we describe in the next section. In this section, we briefly outline the steps in implementing our framework.

The first step is to run an efficient implementation of the Apriori algorithm: We have selected to use the application developed by C. Borgelt which is available on the internet [4] and is well documented.

To generate both the frequent itemsets and association rules, it is required to run Borgelt's application twice because this particular application is capable of

generating either the frequent itemsets or the association rules. One important issue that we paid attention to was using the right options while computing the support levels of the association rules. The default computation of the support level in Borgelt's application is different from the original definition by Agrawal and Srikant [2]. The option "`-o`" is included in command line to adhere to the original definition which we have defined in the Introduction section.

The second step is to translate the results of the Apriori algorithm into a graph specification. In our case study, we carried out this step by converting the support and confidence levels to corresponding node diameters and colors in a spreadsheet.

We then created the graph objects based on the calculated specifications in the previous step. There are a multitude of tools available on the internet for graph visualization [19]. We have selected the `yEd` Graph Editor [24] for drawing our initial graph and generating visually interpretable graph layouts. `yEd` implements several types of graph drawing algorithms including those that create hierarchical, organic, orthogonal, and circular layouts. For interested readers, the detailed information on the algorithms and explanations of the various settings can be found under the program's Help menu.

The final step is to run the available graph layout algorithms and try to visually discover interesting and actionable insights. Our case study revealed that different layout algorithms should be selected depending on whether one is visualizing frequent itemsets or association rules, and on the underlying purpose of the analysis e.g. catalog design, shelf layout, and promotional pricing.

5 Case Study: Market Basket Analysis

The benchmark data set used in this study is provided at [10] and initially analyzed in [5] for assortment planning purposes. The data set is composed of 88,163 transactions and was collected at a Belgian supermarket. There are 16,470 unique items in it. The data set lists only the composition of transactions; different visits by the same customer cannot be identified and the monetary value of the transactions are omitted.

In this section, we present our findings through visual analysis of frequent itemsets and association rules. Frequent itemset graphs generated by `yEd` provided us with guidelines for catalog design and supermarket shelf design. Association rule graph supplied us with inherent relationships between the items and enabled development of promotional pricing strategies.

5.1 Visualizing Frequent Itemsets

Figure 1 depicts the results of the Apriori algorithm that generated the frequent itemsets at support level of 2% for the data set. This graph was drawn by selecting the Classic Organic Layout in `yEd`. Items that belong to similar frequent itemsets and have high support levels are placed in close proximity of each other. From Figure 1, one can easily notice the Items 39, 48, 32, 41, and 38 have very

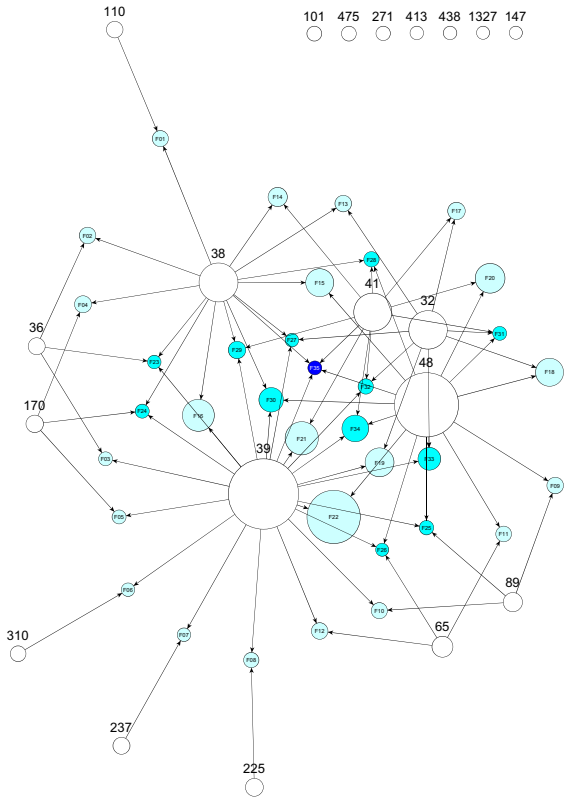

Fig. 1. Visualization of the frequent itemsets through a Classic Organic Layout

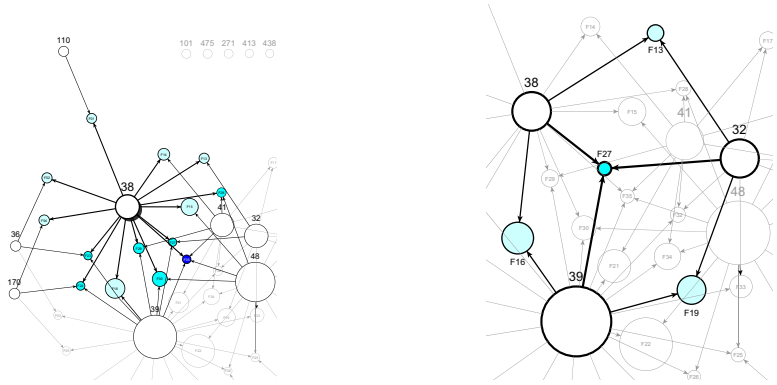

Fig. 2. Querying the visualization of the frequent itemsets (a) Selecting a single item. (b) Selecting three items together.

high support levels, and form frequent itemsets (with up to four elements) among themselves. On the upper corner of the figure, there is a set of items which meet the minimum support level condition, but have no significant associations with any other item.

Besides identifying the most significant and interdependent items, one can also see items that are independent from all but one of the items. For example, Items 310, 237, and 225 are independent from all items but Item 39. This phenomenon enables us to visually cluster (group) the items into sets that are fairly independent of each other. In the context of retailing, catalog design requires selecting sets of items that would be displayed on each page of a catalog. From Figure 1, one can easily determine that Items 310, 237, and 225 could be on the same catalog page as Item 39. One might suggest putting Items 39, 48, 32, 41, and 38 on the same page since these form frequent itemsets with high confidence levels. However, this would result in a catalog with only one page of popular items and many pages of much less popular items. We suggest that the popular items be placed on different pages so that they serve as *attractors*, drawing attention to less popular items related to them.

Another type of insight that can be derived from Figure 1 is the identification of items which form frequent itemsets with the same item(s) but do not form any frequent itemsets with each other. For example, Items 65 and 89 each independently form frequent itemsets with Items 39 and 48, but do not form frequent itemsets with one another. These items may be substitute items and their relationship deserves further investigation.

The frequent itemset visualization can be enhanced by incorporating interactive visual querying capabilities. One such capability could be that once an item is selected, the items and the frequent itemsets associated with it are highlighted. This concept is illustrated in Figure 2 (a) where Item 38 is assumed to be selected. It can be seen that Item 38 plays a very influential role since it forms frequent itemsets with seven of the 13 items. When two or more items are selected, only the frequent itemsets associated with all of them could be highlighted. Thus selecting more than one item could serve as a query with an AND operator. This concept is illustrated in Figure 2 (b) where Items 39, 32 and 38 are assumed to be selected. One can notice in this figure that even though all the three items have high support levels each (as reflected by their large sizes) the frequent itemset F27 (in the middle) that consists of all the three items has a very low support level. When analyzed in more detail it can be seen that this is mainly due to the low support level of the frequent itemset F13 which consists of Items 32 and 38. This suggests that the association between Items 32 and 38 is significantly low, and these items can be placed into separate clusters.

When experimenting with various graph layout algorithms within `yEd` we found the Interactive Hierarchical Layout particularly helpful. The obtained visualization is given in Figure 3, where items are sorted such that the edges have minimal crossings and related items are positioned in close proximity to each other. This visualization can be used directly in planning the supermarket shelf layouts. For example assuming that we would place these items into a single

aisle in the supermarket and that all the items have same unit volume, we can lay out the items according to their positions in Figure 3, allocating shelf space proportional to their node sizes (support levels).

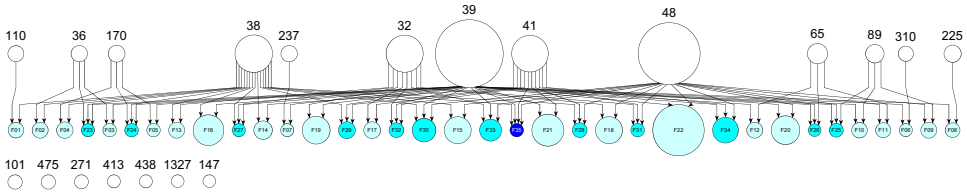

Fig. 3. Visualization of the frequent itemsets through an Interactive Hierarchical Layout

Of course in a real world setting there would be many aisles and the items would not have the same unit volume. For adopting our approach to the former situation we can pursue the following steps: We start with the analysis of a Classic Organic Layout as in Figure 1, and determine from this graph groups of items that will go into aisles together. When forming the groups we try to make sure that the total area of the items in each group is roughly the same. Once the groups are determined we can generate and analyze an Interactive Hierarchical Layout as in Figure 3 for each group and then decide on the shelf layouts at each aisle. Incorporating the situation where the items have significantly varying unit volumes is a more challenging task, since it requires consideration of these volumes in addition to consideration of the support levels.

5.2 Visualizing Association Rules

Figure 4 depicts the results of the Apriori algorithm that generated the association rules at support level of 2% and confidence level of 20% for the data set. This graph was drawn by selecting the Classic Hierarchical Layout in `yEd`.

The figure shows which items are "drivers" that push the sales of other items. For example one can observe at the top of the figure that the rules A01 (Item 110 ⇒ Item 38), A02 (Item 36 ⇒ Item 38), and A04 (Item 170 ⇒ Item 38) all have high confidence levels, as reflected by their red colors. This observation related with high confidence levels can be verified by seeing that the node sizes of the rules A01, A02, and A04 are almost same as the node sizes of Items 110, 36, and 170. For increasing the sales of Item 38 in a retail setting we could use the insights that we gained from Figure 4. Initiating a promotional campaign for any combination of Items 110, 36, or 170 and placing these items next to Item 38 could boost sales for Item 38. This type of a campaign should especially be considered if the unit profits of the driver items (which reside in the antecedent

of the association rule) are lower than the unit profit of the item whose sale is to be boosted (which resides in the antecedent of the association rule).

6 Conclusions and Future Work

We have introduced a novel framework for knowledge discovery from association mining results. We demonstrated the applicability of our framework through a market basket analysis case study where we visualize the frequent itemsets and binary association rules derived from transactional sales data of a Belgian supermarket.

Ideally, the steps in our framework should be carried out automatically by a single integrated program or at least from within a single modelling and analysis environment that readily communicates with the association mining and graph visualization software. Such a software does not currently exist.

In the retail industry new products emerge and consumer preferences change constantly. Thus one would be interested in laying the foundation of an analysis framework that can fit to the dynamic nature of retailing data. Our framework can be adapted for analysis of frequent itemsets and association rules over time by incorporating latest research on evolving graphs [9].

Another avenue of future research is testing other graph visualization algorithms and software [19] and investigating whether new insights can be discovered by their application. For example, Pajek graph visualization software [20] enables the mapping of attributes to line thickness, which is not possible in `yEd`. The visualizations that we presented in this paper were all 2D. It is an interesting research question to determine whether exploring association rules in 3D can enable new styles of analysis and new types of insights.

As a final word, we conclude by remarking that visualization of association mining results in particular, and data mining results in general is a promising area of future research. Educational, research, government and business institutions can benefit significantly from the symbiosis of data mining and information visualization disciplines.

References

1. `http://www.advizorsolutions.com`
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. Proceedings of the 20th VLDB Conference,Santiago, Chile. (1994) 487–499
3. Battista, G. D., Eades, P., Tamassia, R., Tollis, I. G.: Graph Drawing: Algorithms for the Visualization of Graphs. Prentice Hall PTR.(1998)
4. `http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html`
5. Brijs, T., Swinnen, G., Vanhoof, K., Wets, G.: The use of association rules for product assortment decisions: a case study. Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, San Diego (USA), (1999) 254–260.

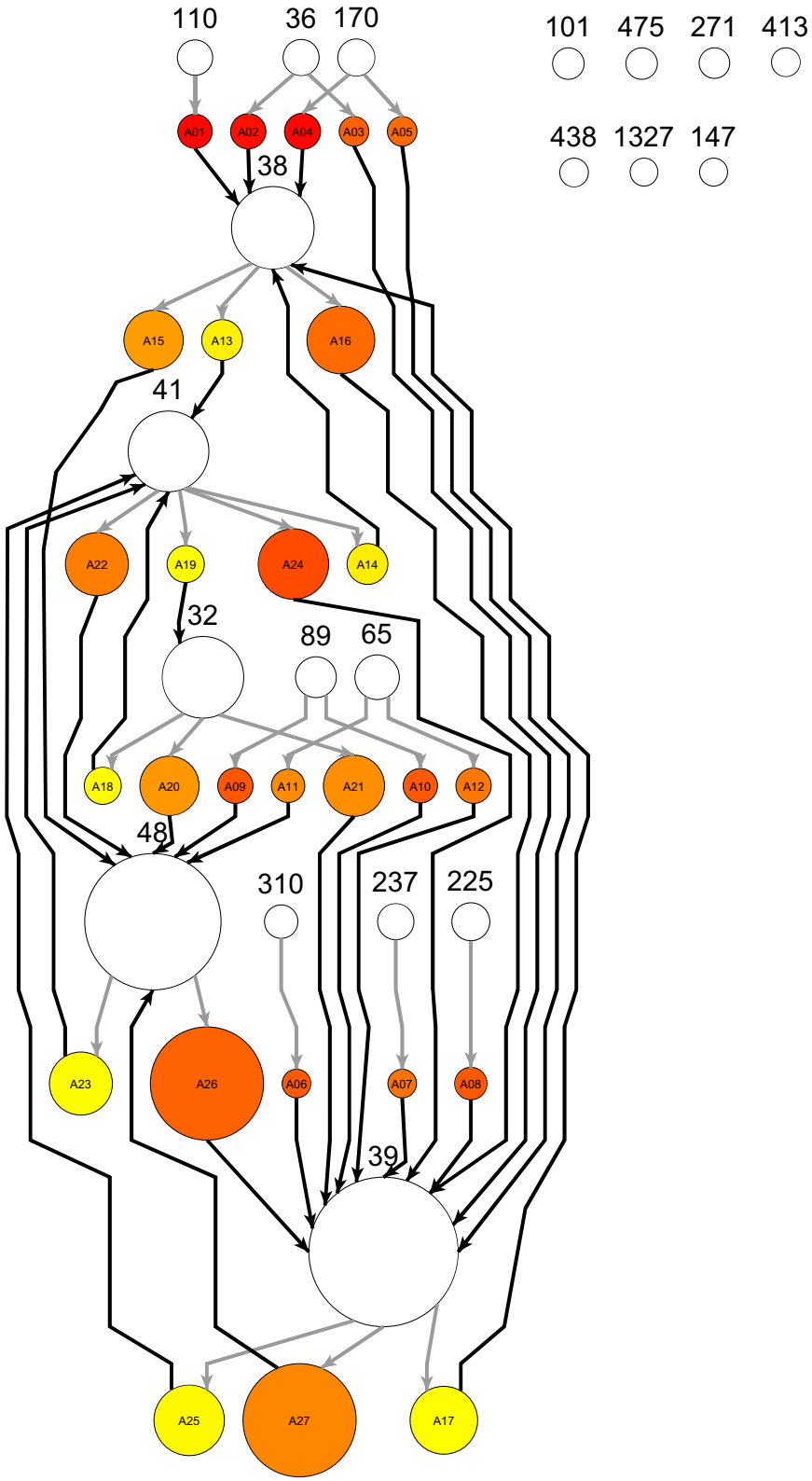

Fig. 4. Visualization of the association rules through an Interactive Hierarchical Layout

6. de Oliveira, M. C. F., Levkowitz, H.: From visual data exploration to visual data mining: a survey. IEEE Transactions on Visualization and Computer Graphics **9**, no.3 (2003) 378–394
7. Demiriz, A.: Enhancing product recommender systems on sparse binary data. Journal of Data Mining and Knowledge Discovery. **9**, no.2 (2004) 378–394
8. Eick, S. G.: Visual discovery and analysis. IEEE Transactions on Visualization and Computer Graphics **6**, no.1 (2000) 44–58
9. Erten, D. A., Harding, P. J., Kobourov, S. G., Wampler, K., Yee, G.: GraphAEL: Graph animations with evolving layouts. Lecture Notes in Computer Science. **2913**, (2004) 98–110
10. `http://fimi.cs.helsinki.fi/data/`
11. Han, J., Kamber, M.: Data Mining Concepts and Techniques. Morgan Kaufman Publishers. (2001)
12. Herman, I., Melançon, G., Marshall, M. S.: Graph visualization and navigation in information visualization: A survey. IEEE Transactions on Visualization and Computer Graphics. **6**, no.1 (2000) 24–43
13. Hoffman, P. E., Grinstein, G. G.: A survey of visualizations for high-dimensional data mining. Chapter 2, Information visualization in data mining and knowledge discovery, Eds: Fayyad, U., Grinstein, G. G., Wierse, A. (2002) 47–82
14. Hofmann, H., Siebes, A. P. J. M., Wilhelm, A. F. X.: Visualizing association rules with interactive mosaic plots. Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining. (2000) 227–235
15. `http://www-306.ibm.com/software/data/iminer/visualization/`
16. Keim, D. A.: Information visualization and visual data mining. IEEE Transactions on Visualization and Computer Graphics. **8**, no.1 (2002) 1–8
17. Kopanakis, I., Theodoulidis, B.: Visual data mining modeling techniques for the visualization of mining outcomes. Journal of Visual Languages and Computing. **14** (2003) 543–589
18. `http://www.miner3d.com/`
19. `http://www.netvis.org/resources.php`
20. `http://vlado.fmf.uni-lj.si/pub/networks/pajek/`
21. `http://www.spotfire.com/`
22. Tan, P., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. Proceedings of SIGKDD. (2002) 32–41
23. Wong, P. C., Whitney, P., Thomas, J.: Visualizing association rules for text mining. Proceedings of the 1999 IEEE Symposium on Information Visualization. (1999)
24. `http://www.yworks.com/`