Market Basket Analysis Visualization on a Spherical Surface

Ming C. Hao, Meichun Hsu, Umeshwar Dayal, Shu Feng Wei, Thomas Sprenger[1], Thomas Holenstein
Software Technology Laboratory
HP Laboratories Palo Alto

HPL-2001-3
December 29th, 2000*

visualization, association, physics-based mass spring, market basket analysis, spherical surface

This paper discusses the visualization of the relationships in e-commerce transactions. To date, many practical research projects have shown the usefulness of a physics-based mass-spring technique to layout data items with close relationships on a graph. We describe a market basket analysis visualization system (MAV) using this technique. This system is described as the following: (1) integrates a physics-based engine into a visual data mining platform; (2) uses a 3D spherical surface to visualize the cluster of related data items; and (3) for large volumes of transactions, uses hidden structures to unclutter the display. Several examples of market basket analysis are also provided.

* Internal Accession Date Only Approved for External Publication

[1] Presently with AdNovum Informatik AG, CH-8005 Zuerich

Market Basket Analysis Visualization On A Spherical Surface

Ming C. Hao, Meichun Hsu, Umeshwar Dayal, Shu Feng Wei
Thomas Sprenger*, Thomas Holenstein
Hewlett Packard Research Laboratories

ABSTRACT

This paper discusses the visualization of the relationships in e-commerce transactions. To date, many practical research projects have shown the usefulness of a physics-based mass-spring technique to layout data items with close relationships on a graph. We describe a market basket analysis visualization system (MAV) using this technique. This system is described as the following: (1) integrates a physics-based engine into a visual data mining platform; (2) uses a 3D spherical surface to visualize the cluster of related data items; and (3) for large volumes of transactions, uses hidden structures to unclutter the display. Several examples of market basket analysis are also provided.

Keywords: visualization, association, physics-based mass spring, market basket analysis, spherical surface

1. INTRODUCTION

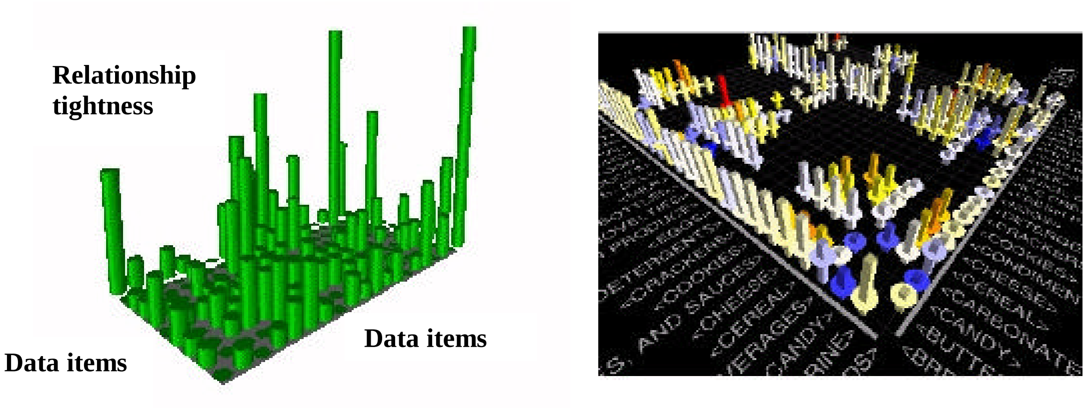
Market basket analysis requires the analysis and mining of large volumes of transaction data for making business decisions. It has become a key success factor in e-commerce. Effective market basket analysis techniques employ association and clustering as methods of analyzing such data. E-commerce transactions often consist of several products (items) that are purchased together. Understanding the relationships across hundreds of product lines and among millions of transactions provides visibility and predictability of product affinity. An example of market basket analysis is that 85% of the people who buy a printer also buy paper at the same time.

There are many technologies [5, 6, 7, 11] that enable the visualization of associations to make business decisions, such as product recommendations, bundling, and cross selling. A common technique for visualizing associations is a SOM [10] map. The SOM technique positions pairs of items on separate axes to visualize their relationships. Figure 1 is the n-matrix representation of the map. The height corresponds to the tightness of the relationships (associations). Comparing the height of the cylinders (association values) enables the easy comparison of the relationships between different pairs of data items. SOM has been employed in the IBM [6], SGI [7], and SAS [11] market basket data mining applications.

For visualizing e-commerce transaction data with a large number of items, we have found that SOM is too restrictive. It only shows a maximum of 10-20 data items. The map becomes cluttered with many units when the number of items grows large. Also, because SOM is a 2-D matrix representation, it cannot use distance to represent the strength of the relationships between products. The following are the recent needs in the new visualization of market basket analysis:

(1) Scale to a large number of products.
(2) Show the products that are directly and indirectly related.
(3) Show the strengths of the relationships (closeness).

*Presently with AdNovum Informatik AG, CH-8005 Zuerich
Correspondence: Email: (ming_hao, mhsu, dayal, tholenst) @hpl.hp.com; sprenger@adnovum.ch



X-axis and Y-axis are the same. Both show data items such as butter, candy, cereal, cheese… Z-axis shows the co-occurrence frequency of a pair of items, such as butter and candy, candy and cheese…

Figure 1: A Conventional Method (SOM) and SGI MineSet Visualization of Association

2. OUR EXPERIMENTAL VISUALIZATION SYSTEM

To analyze a large highly related product space, we are experimenting with a new market basket analysis technique (MAV). The detail method and algorithm of MAV are described in the Hewlett Packard Technical Report [2], in which we describe the integration of a physics-based mass-spring engine and a data mining visualization system [4]. Also, we describe how to cluster related items in a 3D spherical surface for pattern discovery. For scalability, we are experimenting with various methods to aggregate and hide the data.

MAV uses a spherical surface to place items nearby that are frequently purchased together. The "distance" between each pair of items represents the strength of the relationship. The most tightly related item is the item with the highest correlation with other items. These items usually appear in the same shopping basket, e.g. customers usually buy printer and paper together.

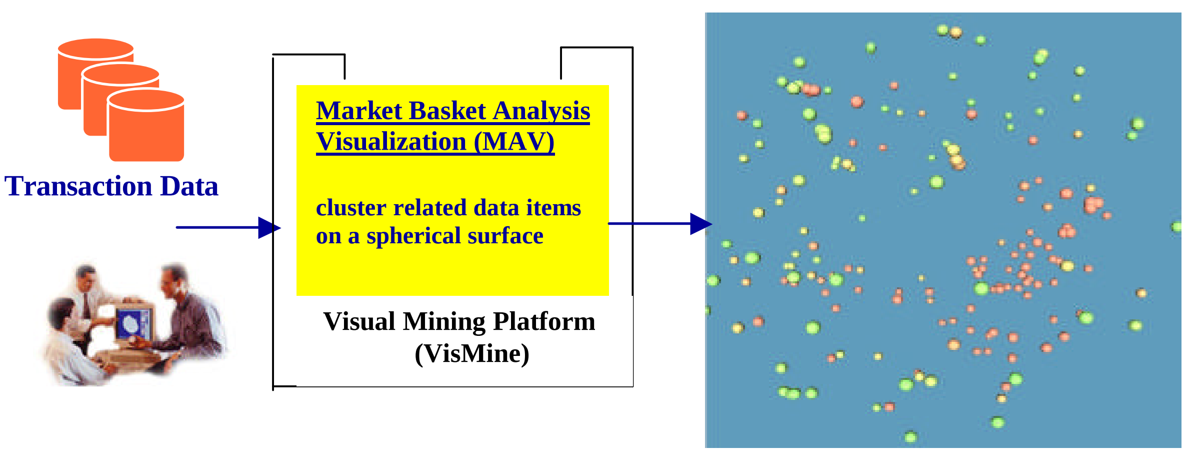


Figure 2: Our Experimental Market Basket Analysis Visualization

3. MARKET BASKET ANALYSIS VISUALIZATION FLOW

The MAV technique is built on Java-based multi-threaded parallelism. MAV processor interacts with the user and data mining engines to select products for visualization. The MAV processor manages the following three different processing states (As illustrated at Figure 3, 4, 5).

3.1. Initial State

First, MAV preprocesses the raw transaction data. MAV prunes and filters irrelevant data. For example, MAV removes data items that have very high or low relationships from the baskets A, B, C etc.

Second, MAV calculates the correlations (association) between pairs of data items. The results are stored in an association matrix. MAV represents items as balls. Initially, items are evenly positioned on a 3D spherical surface. The order of items is based on the association values. The most tightly related item is placed in the center and the others are distributed around it.

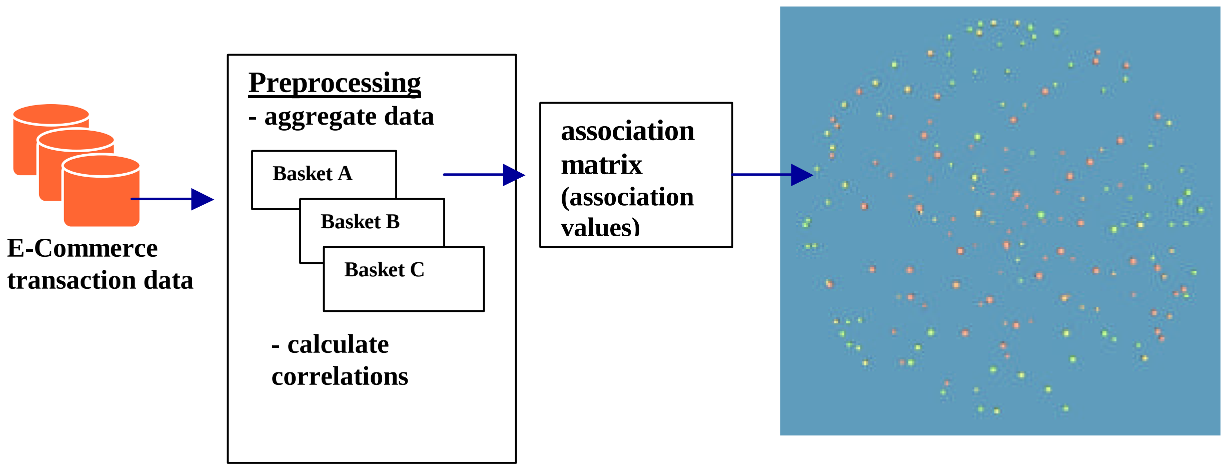


Figure 3: Market Basket Analysis Visualization Initialization

3.2. Relaxation State:

MAV uses a physics-based mass-spring engine [3, 8, 9] to connect product items with springs. The strength of product relationships correlates to the stiffness of the springs between them. The stiffness of the spring is defined in the above association matrix. The mass-spring engine transfers the spring stiffness to the distance between pairs of items. From the principle of a physics-based mass-spring engine, after many iterations, the graph will be relaxed and reach a local minimum. Items with high relationships are automatically moved together and form clusters. Figure 4 illustrates the final graph after the graph has been relaxed for 212 iterations and has reached a local minimum.

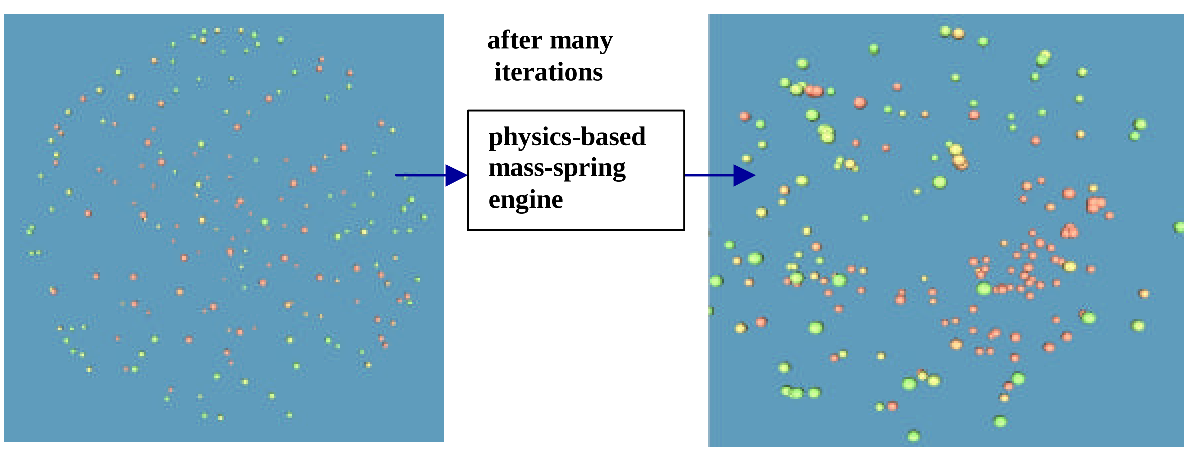


Figure 4: Market Basket Analysis Visualization Relaxation

3.3. Navigation Mining State:

After the graph is relaxed, the user can start navigation to discover and analyze patterns. For example, the user can easily in Figure 5 detect and select cluster 1 to find items for cross selling and up selling. Cluster 1 points to the detailed information about data items, such as directed edges for association directions and colored balls for different product categories. In addition, the user can select an item – C5990A to get the product name, the department name, and the department type.

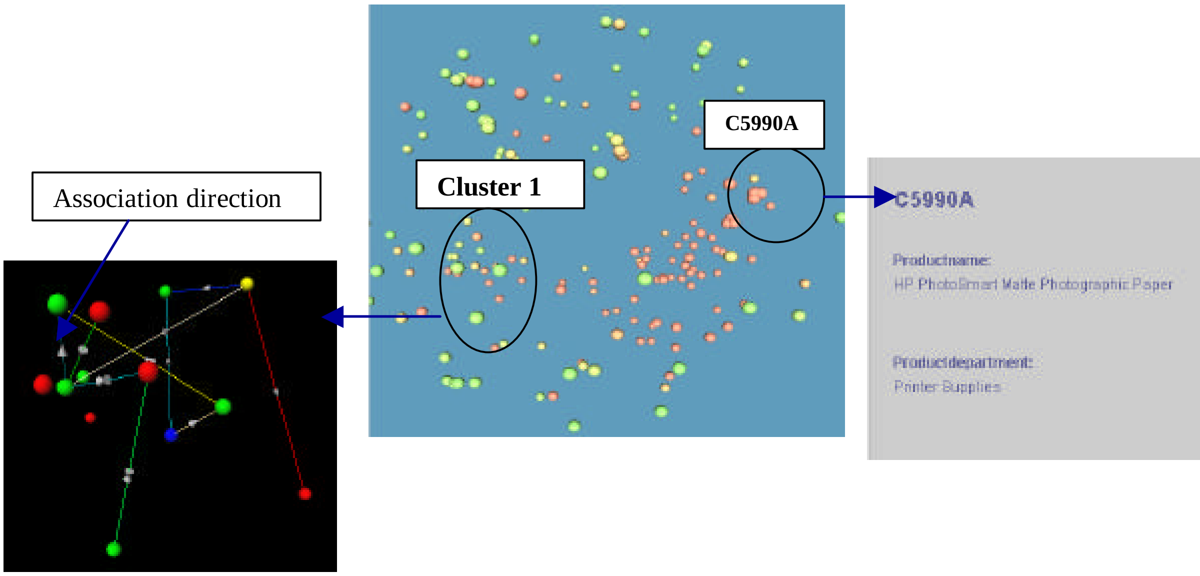


Figure 5: Navigating Market Basket Product Items Relationships

4. APPLICATIONS

The market basket analysis visualization is built on a client-server model. To achieve rapid display, the visualization construction and rendering are done locally in the client site. The data mining and computation execute on the server. The server is integrated with the data mining engine and the data warehouse. In Hewlett Packard Research Laboratories, we have used this market basket analysis visualization technique to experiment with product associations, indirect product associations, user profiling, and other e-commerce applications.

4.1. Product Associations

A business analyst and a manager can navigate a MAV-generated product sales graph and answer questions as to which product groups are frequently bought together, and how strong and in which direction the correlation is. From the previous example in which 85% of the people who buy a printer also buy paper, this visualization may help determine which products should be sold with printers. Also, it helps determine which products may be impacted if the store discontinues selling printers.

Figure 6 illustrates detailed information. MAV allows users to navigate the graph to find out the pattern, such as, customers usually buy the yellow, cyan, and magenta Inkjet cartridges together with Inkjet Paper-24 for printing. Also, not surprisingly customers buy yellow, cyan, and magenta ink cartridges together.

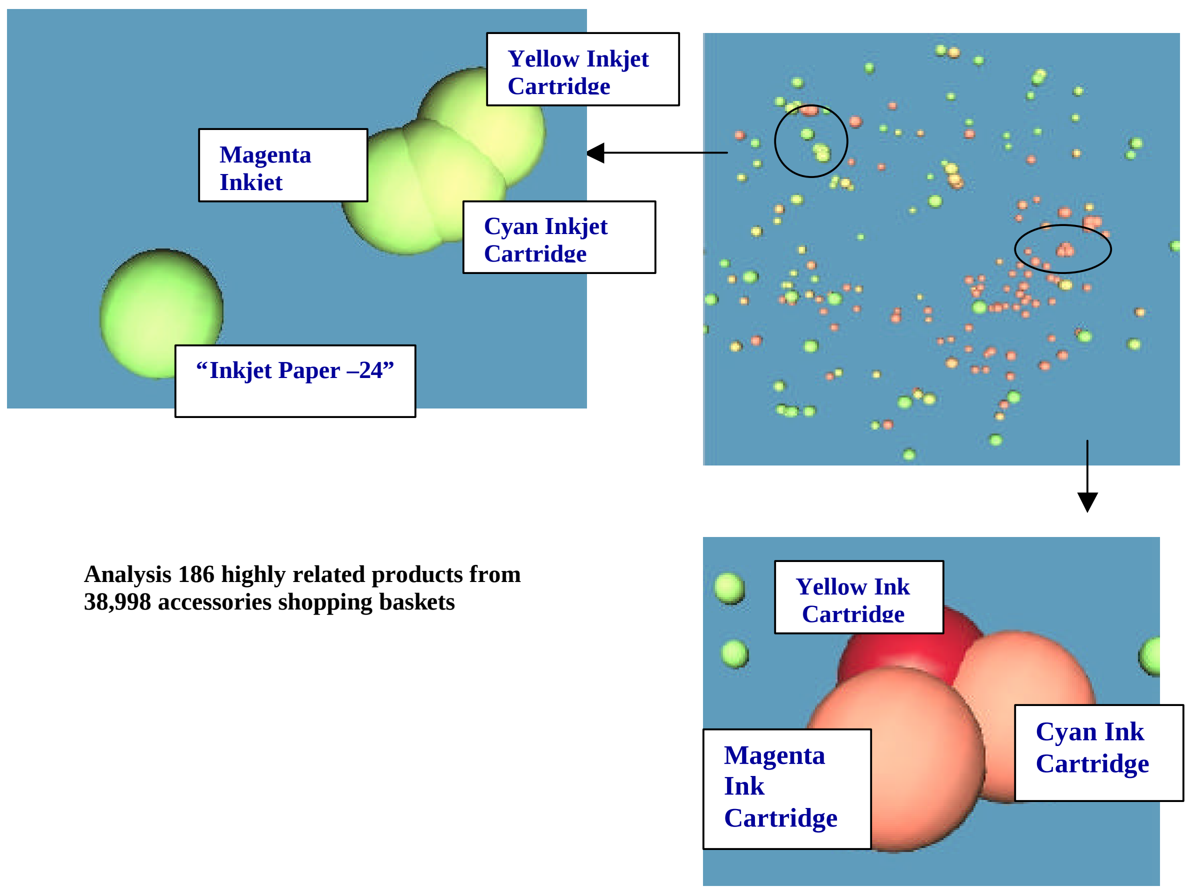
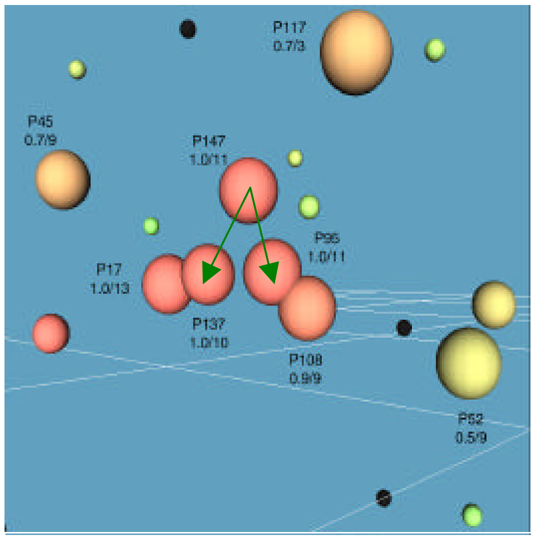


Figure 6: E-Commerce Market Basket Analysis Visualization

4.2. Indirect Association Visualization

The second example shown is indirect product association [1]. The indirect associations are represented by A↔B↔C. If a customer buys item B, and is interested in item A, the store manager may want to recommend C instead of A, if C has a better profit margin, even if there is no relationship between A and C.

As illustrated in Figure 7, P147 is strongly associated with P137. Item P147 is strongly associated with P95, but P137 and P95 are never bought together. For example, this might occur if P147 and P95 are alternative accessories. For business reasons, it may be better for the store manager to recommend that the customer buy P137 when the customer buys P147 not P95.



Arrows: represent the association direction (link) between two products

Link1: from P 147 to P95

Link2: from P147 to P137

Figure 7: An Indirect Association Application Example

4.3. Other Applications

The market basket analysis visualization can be applied to market basket analysis in other business domains such as banking and retail stores. Furthermore, it can be used for other data mining tasks such as user profiling /customer segmentation or text/document clustering.

5. LIMITATION AND SCALABILITY

This system is quite efficient when used with thousands transactions and hundreds products. However, there are millions of e-commerce transaction records involving thousands of products processed daily. In order to visually mine relationships without overwhelming the system and cluttering the display, we are experimenting with the following techniques.

(1) Aggregation and segmentation

MAV dynamically aggregates the items or groups of items as necessary. To avoid processing large volumes of data at the same time, MAV incrementally aggregates data at multiple levels of data mining [9]. The data is partitioned into segments based on the characteristics of the data. Single data items in a transaction are grouped into hierarchical categories.

(2) Thresholding

MAV employs different thresholding techniques to greatly reduce the computation and rendering time. The user is allowed to try different sets of thresholds, such as to display only data items with an association higher than 0.02. Also, MAV eliminates the items with low relationships.

(3) Hiding and focusing

MAV allows the user to focus freely on the products and associations of interest. It supports two levels of focusing policy: cluster-based and item-based. Cluster-based focusing allows the user to view items and their relationships in a selected cluster. Item-based focusing allows the user to view the selected items and their relationships. MAV experiments with various schemes to hide the structures and relationships that are currently not in focus. The hidden items or links appear when the user focuses on them.

(4) Distribution of computation

In order to scale-up, MAV distributes the computation from the rendering onto different processors. For example, the data aggregation and matrix computation can be executed on multiple server sites. The graph layout and rendering are performed on the client sites.

6. CONCLUSION

The market basket analysis using MAV enables a user to mine through different products without being overwhelmed with a large number of market basket transaction data. Our technique still could be enhanced to a large transaction data space. In addition, this technique could be applied to the area of sequential pattern analysis over time.

ACKNOWLEDGEMENT

Thanks to Sharon Beach, Vineet Singh, Craig M. Wittenbrink, Martin Griss from HP Research Laboratories, and to Prof. Markus Gross from Swiss Federal Institute of Technology; for their encouragement and suggestions and to Graham Pollock from Agilent Laboratories for review and comments. Also thanks to Patrick Barthelemy from "Template Graphics Software" for technical support and for allowing us to use their toolkits in our experiments.

REFERENCE

1. Pang-Ning Tan, Vipin Kumar, Jadeep Srivastava, Indirect Association: Mining Higher Order Dependencies in Data (2000). PKDD2000.
2. Ming Hao, Umesh Dayal, Meichun Hsu, Thomas Sprenger, Markus H. Gross, "Visualization of Directed Associations in E-Commerce Transaction Data", HP Laboratories Technical Report, HPL-2000-160.
3. T.C. Sprenger, M.H. Gross, "Ivory – An Object Oriented Framework for Physics-Based Information Visualization in Java", IEEE InfoVis98, North Carolina.
4. Ming Hao, Umesh Dayal, Meichun Hsu, Bob D'eletto, Jim Becker, "A Java-based Visual Mining Infrastructure and Applications", IEEE InfoVis99, CA.
5. Pak Chung Wong, Paul Whitney, Jim Thomas, "Visualizing Association Rules for Text Mining", IEEE InfoVis99, CA.
6 "Quest": IBM Data Mining Technologies.
7. "MineSet": SGI MineSet 3.0 Enterprise Edition.
8. M.H. Gross, T.C. Spenger, J. Finger: "Visualizing Information on a Sphere", IEEE VisInfo97, InforVis 2000, 2000, Utah.
9. T.C. Sprenger, R. Brunella, M.H. Gross, "H-BLOB: A Hierarchical Visual Clustering Method Using Implicit Surfaces", IEEE/VIS2000.
10. SOM: Self Organized Map, http://cmgm.stanford.edu/pbrown/sporulation/additional/websom.html.
11. SAS: Enterprise Miner, http://www.sas.com.