

# Visual Data Mining Techniques for Information Retrieval in Massive Data Sets

<sup>1</sup>Mr. Ravish G K, <sup>2</sup>JC Achutha  
<sup>1,2</sup> Asst. Professor, Dept. of MCA  
The Oxford College of Engineering  
Bommanhalli, Hosur Road, Bangalore-560068  
Email: [gkravish@gmail.com](mailto:gkravish@gmail.com), [achutha.sir@gmail.com](mailto:achutha.sir@gmail.com)

**Abstract:** Visual data mining (VDM) is a collection of interactive methods by which we can explore large data sets by dynamically adjusting parameter values. The main aim of this process is to know the affect of information being presented. Visual data mining finds structure like cluster, unusual observations in large and not necessarily homogeneous data sets based on human perception using graphical methods and user interaction. Visual data mining techniques enable us to make observations without preconception. These techniques are based on computer graphics, visualization metaphors and methods, information and scientific data visualization, visual perception, visual data formatting and 3D environments for information visualization. Also we have to think about the limits of available computational resources.

Keywords: VMD, crowded VDM, Stacked VDM, knowledge discovery

## 1. Visual Data mining Techniques

Visualization has been used routinely in data mining as a presentation tool to generate initial views, navigate data with complicated structures, and convey the results of an analysis. Generally, the analytical methods themselves don't involve visualization. The loosely coupled relationships between visualization and analytical data mining techniques

represent the majority of today's state of the art in visual data mining. The process sandwich strategy, which interlaces analytical processes with graphic visualization, penalizes both procedures with each other's deficiencies and limitations. A stronger visual data mining strategy may lie in tightly coupling the visualizations and analytical processes into one data mining tool. Letting human visualization participate in an analytical process' decision-making remains a major challenge. Certain mathematical steps within an analytical procedure may be substituted by human decisions based on visualization to allow the same analytical procedure to analyze a broader scope of information. There are a large number of visualization techniques which can be used for visualizing the data [6]. Visual data mining tries to integrate visual information retrieval and browsing techniques in one content-based similarity measurement process. Retrieval techniques (extraction of content-based features, vector space model, and distance measure based similarity measurement) are used to induce a basic ordering in the media data. The resulting distance space is presented to the user in a three-dimensional user interface. The user can browse through this space and interact with the visualized media set in a number of ways. Figure 1 visualizes the three-dimensional user interface used for browsing/retrieval. Positive and negative query examples are given as 'x' and 'o'.

The objects are visualized on the plane (X-Y-plane). The floor dimensions are used to display a two-dimensional subspace of distance space. This is sufficient to visualize any relationship of two features in the user interface. Distance of two media objects on the X- and Y-axis corresponds to the distance of their feature vectors [5]. The three-dimensional user interface is the basic building block of the visual data mining user interfaces. It offers rich information on visualized media objects and the retrieval process as well as multiple views on the data.

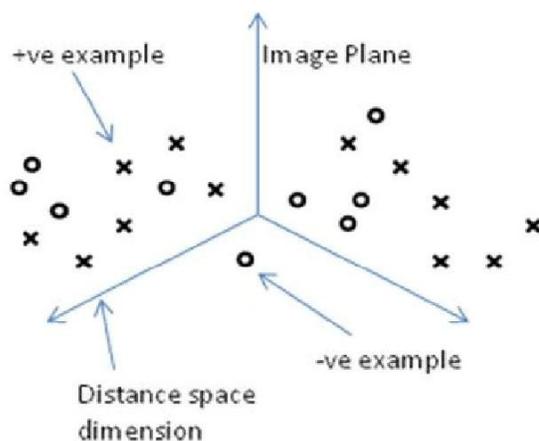


Figure 1 A three dimensional visual data mining user interface

### 1.1 Symbolic VDM Technique

Symbolic VDM Technique[2] involves mapping the values of a various dimensional data item to the features of a symbol. The mapping of attribute values of each data record with the features of symbol creates the generalization. Here the tree dimensions of the figure are used in different ways to figure out the information. Two dimensions are mapped to the display dimensions and the remaining dimensions are mapped to the symbol. If the data items are relatively clustered with respect to the two display

dimensions, the resulting visualization presents texture patterns that vary according to the characteristics of the data and are therefore detectable by pre attentive perception.

### 1.2 Crowded VDM Technique

Each dimensional values is mapped into a colored pixel and grouping them into adjacent areas. It uses one pixel per data value. Crowded VDM technique [2] uses different arrangements for different purposes. The information regarding the hot spots & dependencies can be obtained by different arrangement of the pixels. The recursive pattern technique is based on a generic recursive back and forth arrangement of the pixels and focuses at representing datasets with a natural order according to one attribute. The base element on each recursion level is a pattern of height and width as specified by the user.

### 1.3 Geometric VDM Technique

Geometrical VDM technique [2] finds variations of multidimensional data items. The parallel coordinate technique maps the k-dimensional space onto the two display dimensions by using k equidistant axes which are parallel to one of the display axes. The axes correspond to the dimensions and are linearly scaled from the minimum to the maximum value of the corresponding dimension. Each data item is presented as a polygonal line, intersecting each of the axes at that point which corresponds to the value of the considered dimensions.

### 1.4 Stacked VDM Technique

Stacked VDM techniques [2] are used to display data divided into a top down tree

structure fashion. The data dimensions to be used for dividing the data and constructing the tree structure have to be selected appropriately. Here one coordinate system is embedded into other coordinate system. The display is generated by dividing the outmost level coordinate systems into rectangular cells and within the cells the next two attributes are used to span the second level coordinate system. This process may be repeated one more time. The usefulness of the resulting visualization largely depends on the data distribution of the outer coordinates and therefore the dimensions which are used for defining the outer coordinate system have to be selected carefully. A rule of thumb is to choose the most important dimensions first.

## 2. The future

The field of visual data mining will continue to grow at an even faster pace in the future [4]. In universities and research labs, visual data mining will play a major role in physical and information sciences in the study of even larger and more complex scientific data sets. It will also play an active role in nontechnical disciplines to establish knowledge domains. An advanced form of scatter plot matrix may substitute for the use of covariance and regression in statistics studies. National standards will be developed to govern the functionality and resources of visual data mining. In industries and households across the country, visual data mining will be embedded in public utilities and home appliances. Many searching references such as the yellow pages, dictionaries, and even newspapers will have visual mining capability. The term visual data mining will be included in school text books and literature.

## Conclusion and Future Enhancement:

The Visual data mining has been deployed and aids in knowledge discovery in databases. The major advantages of visual data mining are integration of retrieval and browsing approaches, intuitive visualization of feature data, geographical data and other retrieval-relevant data and easy to handle but still powerful interaction paradigms. In visual data mining the user does not have to deal with giving a number of relevant items he wants to retrieve or with feature weights. Furthermore, he does not have to give textual relevance information during the querying process. Understanding VDM techniques are of benefit to the end users involved with the data mining results and give them more control in the knowledge discovery process.

## References

- [1] A. Inselberg and B. Dimsdale, "Parallel coordinates: A tool for visualizing multi-dimensional geometry," in Proc. Visualization 90, San Francisco, CA, 1990, pp. 361–370.
- [2] Daniel A. Keim, "Information Visualization and Visual Data Mining", IEEE transactions on visualization and computer graphics, Vol 7, No. 1, January-March 2002.
- [3] J. LeBlanc, M. O. Ward, and N. Wittels, "Exploring n-dimensional databases," in Proc. Visualization '90, San Francisco, CA, 1990, pp. 230–239.
- [4] Pak Chung Wong, an Article, Pacific North West National Laboratory, 0272 - 1716/99/IEEE September/October 1999
- [5] Horst Eidenberger, "Visual Data Mining", Vienna University of

Technology, Institute of Software  
Technology and Interactive Systems,  
Favoritenstrasse 9-11, 1040 Vienna,  
Austria, 2004.