Data Clustering and Cluster Mapping or Visualization in Text Processing and Mining

Abstract: The focus of this paper is on a cooperative use of the text data clustering and mapping as visualization-based analysis tools. Whether we expose a generic approach in text processing and mining, we only concentrate on the two-middle steps of the process: data clustering and cluster mapping. In the data clustering analysis step, we use the axial k-means (AKM) algorithm: an iterative partitioning unsupervised winner-take-all (WTA) method, producing overlapping clusters. In the step of mapping the clusters, we use a nonlinear multilayer perceptron (MLP) with two hidden layers. Finally, the map is proposed as an analysis device rather than of visualization. It allows the analyst to evaluate the relative position of clusters which are indicators of themes induced from data themselves.

1 Introduction

It has been claimed that we live in the information age. In this context, competitive pressure in either business or research has led many organizations to use data mining technology. This technology is designed to help decision-makers detect hidden patterns in large volumes of data. Information can take the form of written texts, spoken utterances, factual records and graphical images. In dealing with the textual information type, this paper presents our approach to text mining, which is based on cluster analysis and mapping large volumes of text-data.

The organization of this paper is the following. Section 2 outlines our understanding of text mining and its process. Section 3 describes the text-data clustering step - the AKM algorithm used and its results. Section 4 presents the use of MLP for mapping the clusters that have been obtained by the AKM. The knowledge organization that the clusters designate can be graphically represented. As we shall see the map provides an overview in which cluster positions and relationships can be examined.

2 Text Mining Process Overview

Text mining consists of extracting information from hidden patterns in large textual collections. A very large amount of information is available in textual form in databases and online information sources. In this context, manual analysis and effective extraction of useful information are not possible. It is relevant to provide automatic tools for analyzing large textual collections. The results can be important both for the analysis of the collection, and for providing intelligent navigation and browsing methods (Feldman et al, 1998; Landau et al, 1998).

The text mining process can be organized roughly into five-major steps: [1] Data Selection, [2] Term Extraction and Filtering, [3] Data Clustering, [4] Cluster Mapping or Visualization, [5] Result Interpretation. While performing a particular operation, one often finds it is necessary to revise the operations performed earlier. For example, after displaying the results of a clustering operation, it may be necessary to select additional data, in which case the data selection is repeated. The process is both iterative and interactive, involving steps with many decisions being made by the user or the analyst. In this paper, we focus only on the two-middle steps of the process, the data clustering and the cluster mapping. The former involves the generation of clusters using clustering methods, while the latter consists
of positioning the clusters on a global map in order to display the topical organization of knowledge.

3. Text-Data Clustering

Clustering is used to segment a database into subsets, i.e., clusters. The members of each cluster share a number of interesting properties. The results of a clustering method can be used in one of two ways. First, they can be used for summarizing the contents of the target database by considering the characteristics of each cluster created, rather than those of each record in the database. Secondly, they can be used as input to other methods, e.g. supervised induction. Here, we use the results of our clustering method in the first of these two ways. Clusters are a smaller and more manageable data set. In our application, the target database is 1834 patents about transgenic plant technology indexed by 724 keywords, and recorded in the period 1978-1997. Thus the data matrix is \( X(n, p) \) where \( n \) is equal to 1834 and \( p \) to 724.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>43.04</td>
<td>180</td>
<td>202</td>
<td>164</td>
<td>22</td>
<td>166</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>38.52</td>
<td>180</td>
<td>192</td>
<td>184</td>
<td>35</td>
<td>115</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>34.86</td>
<td>179</td>
<td>199</td>
<td>172</td>
<td>43</td>
<td>157</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>34</td>
<td>196</td>
<td>200</td>
<td>180</td>
<td>29</td>
<td>163</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>31.24</td>
<td>143</td>
<td>124</td>
<td>133</td>
<td>26</td>
<td>146</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>27.2</td>
<td>143</td>
<td>155</td>
<td>138</td>
<td>34</td>
<td>123</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>26.2</td>
<td>104</td>
<td>116</td>
<td>82</td>
<td>31</td>
<td>90</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>25.32</td>
<td>117</td>
<td>121</td>
<td>113</td>
<td>26</td>
<td>101</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>24.31</td>
<td>106</td>
<td>133</td>
<td>104</td>
<td>53</td>
<td>65</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>21.31</td>
<td>83</td>
<td>85</td>
<td>73</td>
<td>40</td>
<td>65</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>19.44</td>
<td>103</td>
<td>108</td>
<td>106</td>
<td>48</td>
<td>90</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>18.75</td>
<td>76</td>
<td>78</td>
<td>73</td>
<td>32</td>
<td>42</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>12.58</td>
<td>50</td>
<td>50</td>
<td>46</td>
<td>33</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>11.55</td>
<td>68</td>
<td>86</td>
<td>68</td>
<td>66</td>
<td>59</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>7.4</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>28</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>5.76</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>39</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>4.90</td>
<td>26</td>
<td>28</td>
<td>26</td>
<td>74</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>4.06</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>48</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>1.55</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>31</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1.34</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1. Statistical Description of the Clusters produced by the AKM Algorithm as follows: [1] Inertia, [2] Number of documents constructing the class, [3] Number of documents whose coordinates are greater than the threshold, [4] Number of specific documents, [5] Number of keywords defining the theme, [6] Number of authors, [7] Number of information sources, it is here empty.

3.1 Clustering versus Classification

Often classification means assigning new data to one of an existing set of possible classes. We are concerned with the problem of automatic discovery of classes in data (called clustering, or unsupervised learning), rather than with the generation of classes from labeled examples (called supervised learning). In data mining, classification is a learning function that classifies data into one of several predefined classes. Clustering is a descriptive task where one seeks to identify a finite set of categories or clusters to describe the data (Fayyad et al, 1996, p. 13, 14). As used in this paper, clustering means finding the classes themselves from a given set of data. Unlike the so-called supervised methods of classification, where the algorithm learns how to distinguish user-specified classes within data, unsupervised classification, or clustering, consists of identifying the statistically significant classes within data themselves.

3.2 Axial K-Means Model Overview

In the data clustering-analysis step, we have used an AKM algorithm. It is an iterative partitioning unsupervised WTA method. Unlike computing the centroids of the clusters, this algorithm computes the clusters in the form of axes on which data and their attributes are projected. In our case, the data are patent records, and their knowledge content is expressed
by keywords or index terms. The algorithm can be parameterized with the maximum number of clusters desired and the threshold of co-ordinates of the documents and keywords on the axes, producing clusters of a particular type. These clusters overlap because a document or a keyword can belong to several clusters at once; The documents or keywords are ranked according to their degree of resemblance to the cluster prototype.

As its name suggests, the axial k-means method is a variant of the k-means method developed by MacQueen (1967). The AKM model is a synthesis of factor analysis and cluster analysis, drawing on the neural network formalism of Kohonen's model - the associative projection model (Lelu, 1993). The algorithm considers the collection of bibliographic references as a cloud of points in a multidimensional space where each dimension corresponds to a keyword. It represents clusters as vectors pointing towards areas of highest density. The AKM algorithm determines the $k$ clusters as the $k$ axes passing through the origin of the space, or the $k$ unit vectors pointing in the direction of the axes.

The position of the $k$ axes is initialized either at random or by the first $k$ documents. Then the square projections of each normalized document are calculated upon the $k$ axes, computing the scalar products of the normalized document with the unit vectors of the $k$ axes. Each document is allocated to the cluster $k$ where the projection on the axis is maximal. To take this allocation into account in the adaptive form of the algorithm, the position of the axis is immediately recalculated. In the iterative form it is recalculated after all the documents have been treated. By successive iterations, the axes are repositioned and they are stabilized in the high density areas of the data point cloud. The result is a strict partition of the documents.

To obtain overlapping clusters, a "typicality threshold" is defined. While a document belongs to the cluster with which it was associated during the final pass, it can also belong to a different cluster if the value of its projection upon this second axis is greater than the threshold. A document may thus belong to several clusters if the values of its projections upon the corresponding axes are greater than the threshold. Documents belonging to a given cluster can be ranked according to the value of their projection upon the axis that represents the cluster. This order corresponds to a decreasing order around a central member or prototype of the cluster which is a fictitious document positioned exactly on the axis. Using the values of the components of the cluster unit vector, the AKM algorithm defines a partition of keywords from the documentary corpus in the same way. As for the documents, the keyword partition can lead to overlapping clusters. A keyword can belong to several clusters and keywords are ordered according to the decreasing pertinence to the cluster prototype. The weighting that is used to determine the value of pertinence allows the bringing out of specific (or typical) keywords for the cluster. These keywords are those frequent in the cluster and rare in the document collection overall. The name given to each cluster by the system is the keyword which best represents the group of documents in that cluster.

This "overlapping" technique of clustering is an alternative to "all or anyone" classification practices. It is based on Rosch's critique of classical views about the cognitive ways to form categories and to group items together in any defined category. The idea is that the categories are built around a central member or prototype. This is a representative example which shares the most features with other members of the category, while sharing either few or not any feature with elements drawn from outside the class (Gardner, 1987). In the field of the artificial intelligence, the theory of the prototypes and "typicality" has been adopted for implementing a model of reasoning that is known as classification-based reasoning (Napoli, 1991).

Knowledge can be organized following two goals. Either it facilitates the access to that which is of the interest for the user (seeker) or facilitates the analysis of it. This last goal is the
objective that we intend when we use cluster analysis. Note that the attempt of cluster analysis is to organize a data set into homogenous groups of cases or entities called clusters and thus helping its analysis. As we have already said in earlier papers, the information analysis which is based on cluster analysis involves the use of three types or levels of indicators: **Keywords** as content indicators of the knowledge conveyed by the documents; **Clusters** as indicators of the topics or the centers of interest contained around which the searched data aggregates; and **Maps** as strategic indicators of the relative position of the topics in the knowledge space covered by the documents analyzed. We turn now to the question of representing in a two dimensional plane ($R^2$) the results of the application of the AKM algorithm: the cluster matrix $X(m, p)$ where $m$ is equal to 20 clusters and $p$ to 724 keywords.

### 3.3 From Clusters to Mapping

Once the clusters were analyzed (and also renamed if deemed necessary by analysts), the 20 clusters were then organized into two-major categories. The first category covers "applications" and comprises 35% of the clusters containing 24% of the patents. The second category covers "methods and techniques", and is made up of 40% of the clusters and 49% of the patents. The remaining clusters (25%) and patents (27%) make up a third category covering miscellaneous aspects. This information which describes the thematic organization of clusters, each one indicating a topic (a theme), can be displayed in standard graphical figures. An alternative to standard representation figures as a way of analyzing the complexity of the knowledge organization is the procedure of mapping clusters. Now the problem is how to project the multidimensional cluster matrix $X(m, p)$ into a two-dimensional map. The analyst will use both the clusters and the map as indicators as we have argued. So the task will be to construct a spatial representation that can facilitate the understanding of knowledge organization such as is extracted from data by means of a clustering process.

### 4. Visualizing Information

The objectives of this section are twofold: first, to provide a brief description of the multi-layered network that we have used; and second, to illustrate its application as a device for the mapping of clusters obtained by the AKM method.

#### 4.1. Using Nonlinear Multilayer Perceptron

We turn here to the question of the capabilities of a network having layers of weights and hidden units of mapping clusters. Such networks are generally called multilayer perceptron (MLP). Interested readers may refer to Bishop (1998) for more extensive discussion. We use the MLP for its "spatial projection properties" since they can be used to map the clusters.

The nodes or units in the network are organized into input, hidden, and output layers. There is a separate connection from each input node to each hidden node and from each hidden node to each output node. Each connection has a connection strength or weight. Each node in the hidden or output layers receives inputs from other nodes in the preceding layer. The network architecture corresponds with the task that the MLP has to, perform. This architecture is presented in figure 1. The input and output layers have an equal $p$ number of nodes. The first hidden layer has 10 nodes, and the second hidden layer has only two nodes. This MLP takes as input vectors of the size $p$ and firstly compresses them into 10 dimensional vectors, and then into 2 dimensional vectors before restoring the information at the output. Here $m$ denotes the number of clusters (points), and $p$ the number of keywords (dimensions). The activation value of the hidden and output nodes is computed by applying a nonlinear activation function. It is sigmoidal: $f(x) = 1/(1 + \exp(-x))$. 


The MLP is a feed forward-layered network, and its learning procedure, and is the back propagation of error using gradient descent. The network learns by adjusting the interconnection weights between nodes. The learning process is a self-association supervised learning process. The network is trained to adjust its output as close as possible to the input. For each input, the individual squared error (ISE) corresponds to the Euclidean distance between the input and output vectors. The back-propagation learning algorithm uses gradient descent according to the Widrow and Hoff rule. The learning level of the network is known due to the mean squared error (MSE) which is calculated periodically. The MSE is given by the following formula: \[ \text{MSE} = \frac{1}{m} \sum ISE \] where for a data \( x \), \( ISE = y \cdot x^2 \).

The learning process is stochastic; the network connections are corrected after processing each datum \( x \). The connections of the two nodes of the second hidden layer are calculated simultaneously. Once the convergence has been achieved, the outputs of the two nodes of the second hidden layer define the coordinates on the map for each cluster. The data (clusters) are passed through the MLP without any transformation. This implies that the network contains a polariser node. It is a node whose output is always equal to 1. The polariser node is also designated by the term bias and noted \( x_0 (x_0 = 1) \).

**Figure 1**
A three-layer perceptron with \( p \) value inputs and \( p \) outputs and two layers of hidden units. First hidden layer: \( x_i = f(\sum [i=1, p] w_{ij}x_j) \). Second hidden layer: \( x_k = f(\sum [k=1, 10] w_{jk}x_j) \). Output layer: \( y_1 = f(\sum [1=1, 2] w_{ik}x_k) \). In the formulas, \( x_j \) and \( x_k \) are the outputs of nodes in the first and second hidden layers; \( w_{ij} \) is the weighting matrix of the connections from the input to the first hidden layer, and \( w_{jk} \) and \( w_{ik} \) are the weighting matrix of the connections between the first and second hidden layers, and between the second and the output layers respectively.

**Figure 2**
MLP Map of the 20 clusters. For the titles of clusters level 1: [13-16-17] and [1-2-6-9-18] but they are related at RCA levels 2 and 3. For other clusters and statistical patterns see table 1. Two groups of clusters can be observed as RCA isolated.

### 4.2. Mapping the Clusters
Once the database is clustered, using a visualization component (the map), the analyst could examine the generated clusters to establish which ones are useful or interesting. On the other hand, knowledge organization necessitates the use of interactive visualization techniques that allow the user to quickly and easily explore the type of information displayed. The map will provide an overview in which cluster relationships can be examined.
Appropriate display of clusters can give the analyst an insight that can be difficult to get from reading tables of output or simple summary statistics, such as found in table 1.

Figure 2 shows the map. This map represents an analysis tool insofar as it allows the analyst to evaluate the relative position of clusters. In order to facilitate the readability of the map, we applied the related component analysis (RCA) as we explained in an earlier paper (Polanco et al., 1998). This method is based on graph theory, and defines the related components that represent the relative closeness between clusters. Levels are calculated from the distances between clusters. The highest level is defined by the minimum distance between clusters and the lowest by the maximum distance between clusters. At a given level, two clusters are connected if their distance is lower than the maximum threshold of that level. Once the connections are calculated, sets of clusters are defined as those which are linked by a connection path. This operation is repeated for each level. Allowing the cluster configurations to be visualized, this technique facilitates the interpretation of the map. This technique also gives the analyst the means of verification if maps respect the distances between the clusters, and therefore the concentration of some clusters and the isolation of others.

The map shows us the knowledge organization that is represented by clusters (themes) that appear well linked on the bottom. A first group formed by 01 (plasmids), 09 (resistance), 18 (transformation), 06 (infection resistance), 02 (protein), and a second set: 13 (agrobacterium), 16 (transgenesis methods), 17 (in vitro cultures). The related components of the three highest levels are shown on this figure. The isolated clusters (themes) around the map do not present significant links with any other cluster. This is the global knowledge organization that the system provides to expert analysts. A hypertext interface makes possible the exploration of this knowledge organization interactively.

5. Conclusions

We have attempted to present our approach to text mining, which is based on AKM cluster analysis and nonlinear MLP for mapping the clusters. An important assumption that is made throughout the article is that the knowledge organization that the clusters designate can be graphically represented. In fact, the map provides an overview in which cluster positions and relationships can be examined. We argue that the "overlapping" technique of clustering is an alternative to "all or anyone" classification practices. On this point, note that the AKM model is based on the same theory of the prototypes and "typicality" which has been adopted for implementing a model of reasoning known as classification-based reasoning. We also argue that the applied MLP is a nonlinear alternative to standard linear principal component analysis (ACP) for mapping.

Now we will turn to the self-organizing map (SOM, see Kohonen, 1997) and compare with the approach that we come to present. The SOM deals with the most popular network algorithm of the unsupervised learning category for clustering and mapping knowledge from text data (Chen et al., 1998).

Acknowledgements

The authors would like to thank Professor Patrick Gallinari for his useful comments and suggestions regarding the MLP. The MLP development has been carried out with the help of Fabien Manonviller, software engineer student at Université Blaise Pascal (Clermont-Ferrand). His collaboration is gratefully acknowledged.

References


