Cluster-Based Visualization of Concept Associations

Nees Jan van Eck, Flavius Frasincar, and Dwain Chang
Econometrics Institute, Erasmus School of Economics, Erasmus University Rotterdam
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
{*nvaneck,frasincar*}*@few.eur.nl* *279457dc@student.eur.nl*

Abstract

One of the steps in the knowledge domain visualization process is the display of the low-dimensional space in which the items under analysis are positioned. In this paper we investigate the visualization of concept associations by means of concept clustering and edge bundling, two techniques that help alleviate the cluttering caused by concept labels and edges. Hierarchical clustering is used for the identification of the different domain categories (i.e., clusters). The edges between pairs of concepts belonging to two clusters are aggregated in a single edge connecting the clusters' centers. The proposed approach enables us to discover the hierarchical structure of our knowledge domain and analyze the strength between the different domain categories.

1 Introduction

Researchers that want to investigate a certain scientific domain are faced with a large number of documents to be explored. Search engines are only a partial solution to this problem as hundreds of documents are usually returned in response to a query of a user. The manual inspection of these documents is a time-consuming and effort-intensive task. Moreover, due to the dynamics of scientific fields such activities need to be repeated over and over again. The field of knowledge domain visualization (KDViz) [3, 4, 5] aims at alleviating this problem by proposing techniques that support the visual exploration of the knowledge of a scientific domain. These techniques are able to automatically distill from large document corpora the most relevant aspects of a domain and present the extracted information in a visual manner.

Following [3], we divide the process of KDViz into the following six steps: (1) collection of raw data, (2) selection of the type of item to analyze, (3) extraction of relevant information from the raw data, (4) calculation of similarities between items based on the extracted information, (5) positioning of items in a low-dimensional space based on the similarities, and (6) visualization of the low-dimensional space.

In this paper the main focus is on the last step of the KDViz process, i.e., the visualization of the low-dimensional space in which the items under analysis are positioned. In previous work [11] we have used concept maps that present associations between concepts in a scientific field. In such maps, the stronger the association between two concepts, the smaller the distance between them. A shortcoming of the used visualization method is that labels used for identifying concepts have the tendency to overlap each other.

One way to improve the visualization of concept maps is to combine it with the presentation of concept density estimates [8]. Based on the concept density maps we were able to detect concept clusters (i.e., regions of high density), which represent areas of strongly associated concepts. The concepts on top of each region give an indication of the topic of the cluster. One of the deficiencies of this method is that clustering is based on the user's visual interpretation of the map.

In this paper we look at an alternative way of clustering concepts, which allows the automatic computing of clusters and their content. We choose to use hierarchical clustering [7] because one does not have to specify the number of clusters in advance and because it returns a hierarchy of clusters instead of a set of flat clusters, as is the case for most other clustering methods. The proposed clustering enables us to understand the hierarchical structure of the domain knowledge under analysis.

A map of concepts can be shown as an undirected graph, where an edge is generated if the association between two concepts is above a certain threshold value. Drawing these edges as straight lines usually produces additional cluttering (besides the cluttering due to concept labeling). In addition to concept clustering, in this paper we also investigate the bundling of edges [1, 6] between concepts corresponding to two clusters, as a way to eliminate the visual cluttering associated with edge drawing.

The rest of the paper is structured as follows. Our approach for the implementation of the KDViz process is discussed in Section 2. The proposed cluster-based visualization approach that fits in the KDViz process is presented in Section 3. Finally, conclusions and future research directions are discussed in Section 4.

2 Knowledge Domain Visualization Process

In this section, we discuss the way in which we implemented each of the KDViz steps identified in the previous section. Our approach, which we successfully applied in our previous work [9], is summarized in Table 1.

2.1 Step 1: Collection of raw data

The first step of the KDViz process is the collection of appropriate data. Since domain visualizations are typically constructed on the basis of a corpus of scientific texts, one has to collect these texts first. In this paper, the raw data consist of a corpus containing abstracts of publications taken from the repository of the Erasmus University Rotterdam[1]. This repository, called RePub, provides access to all electronic publications of researchers affiliated with the Erasmus University Rotterdam. The repository is organized into three main research themes: *Economics & Management*, *Medicine & Health*, and *Law, Culture & Society*. In this paper, we focus on the theme of *Economics & Management*. At the time of collecting the data, the repository contained 1,889 publications for this theme. The abstracts of these publications were retrieved using an RSS feed.

2.2 Step 2: Selection of type of item

The second step of the KDViz process is the selection of the type of item to analyze. The type of item to analyze depends on the question one wants to answer. The most common types of items are journals, articles, authors, and descriptive words or concepts. Each type of item can be used to visualize a different aspect of a scientific field. In the present study, we choose to analyze concepts.

[1] `http://repub.eur.nl` (accessed on February 28, 2007).

Table 1. Summary of the way in which the KDViz process is implemented in this paper.

Step of the KDViz Process	Implementation
(1) Collection of raw data	Abstracts of papers in the field of economics and management
(2) Selection of type of item	Concepts
(3) Extraction of information	Co-occurrences frequencies
(4) Calculation of similarities	Association strength
(5) Positioning of items	VOS
(6) Visualization	Cluster-based visualization

2.3 Step 3: Extraction of information

The third step of the KDViz process is the extraction of relevant information from the raw data collected in the first step. In this paper, the relevant information consists of the co-occurrence frequencies of concepts extracted from the corpus of abstracts taken from the *Economics & Management* section of the RePub repository. The co-occurrence frequency of two concepts is extracted from the corpus of abstracts by counting the number of abstracts in which the two concepts both occur. To identify the concepts that occur in an abstract, one needs a thesaurus of the scientific field with which one is concerned. In the present study, we make use of the OECD Macrothesaurus[2]. The OECD Macrothesaurus is a multilingual thesaurus containing concepts from the field of economics. The concepts in the thesaurus are organized into 19 main categories. We consider 11 out of these 19 categories as relevant to our corpus of abstracts. From these 11 categories, we only take into consideration those concepts that occur in at least four abstracts in the corpus. This is done because we consider the amount of data on concepts occurring in less than four abstracts too limited for a reliable analysis. In total, 252 concepts occur in at least four abstracts of the corpus. For these concepts, the co-occurrence frequencies are counted.

2.4 Step 4: Calculation of similarities

The fourth step of the KDViz process is the calculation of similarities between items based on the information extracted in the third step. Similarities between items are usually calculated by normalizing the co-occurrence frequencies of the items. In this paper, we also take this approach. To normalize the co-occurrence frequencies of the 252 concepts obtained in the previous step, we use the so-called association strength [9]. The aim of this measure is to normalize co-occurrence frequencies in such a way that concepts occurring in many abstracts and concepts occurring in only a few abstracts can be compared in a fair way. The association strength a_{ij} of the concepts i and j is defined as

$$a_{ij} = \frac{mc_{ij}}{c_{ii}c_{jj}} \qquad \text{for } i \neq j, \tag{1}$$

where c_{ij} denotes the number of abstracts in which the concepts i and j both occur, c_{ii} denotes the number of abstracts in which concept i occurs, and m denotes the total number of abstracts.

[2] `http://info.uibk.ac.at/info/oecd-macroth/` (accessed on February 28, 2007; no longer available).

2.5 Step 5: Positioning of items

The fifth step of the KDViz process is the positioning of items in a low-dimensional space based on the similarities calculated in the fourth step. This step is usually performed using a dimensionality reduction technique. These techniques are able to represent multivariate data in a small number of dimensions. In the case of KDViz, this means that high-dimensional item similarities are represented in a two- or three-dimensional space that can be visually interpreted. Multidimensional scaling (MDS) [2] is one of the most commonly used ordination method in the literature on KDViz. However, it is our experience that MDS does not always provide satisfactory results when it is used for KDViz. In this paper, the positioning of the 252 concepts in a low-dimensional space based on their association strengths is therefore accomplished using a method that we called VOS, which is an abbreviation for *visualization of similarities*. We now briefly introduce this method. A more elaborate discussion of VOS, including an analysis of the relationship between VOS and MDS, is provided elsewhere [10].

Let there be n concepts. The aim of VOS is to provide a two-dimensional space in which the concepts $1, \ldots, n$ are located in such a way that the distance between any pair of concepts i and j reflects their association strength a_{ij} as accurately as possible. Concepts that have a high association strength should be located close to each other, whereas concepts that have a low association strength should be located far from each other. The idea of VOS is to minimize a weighted sum of the squared Euclidean distances between all pairs of concepts. The higher the association strength of two concepts, the higher the weight of their squared distance in the summation. To avoid solutions in which all concepts are located at the same coordinates, the constraint is imposed that the sum of all distances must equal some positive constant. In mathematical notation, the objective function to be minimized in VOS is given by

$$E(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i<j} a_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|^2, \qquad (2)$$

where the vector $\mathbf{x}_i = (x_{i1}, x_{i2})$ denotes the location of concept i in a two-dimensional space and $\|\cdot\|$ denotes the Euclidean norm. Minimization of the objective function is performed subject to the constraint

$$\frac{1}{n(n-1)} \sum_{i<j} \|\mathbf{x}_i - \mathbf{x}_j\| = 1. \qquad (3)$$

Note that the distances $\|\mathbf{x}_i - \mathbf{x}_j\|$ in the constraint are not squared. We numerically solve the constrained optimization problem of minimizing (2) subject to (3) in two steps. We first convert the constrained optimization problem into an unconstrained optimization problem. We then solve the latter problem using a majorization algorithm [2]. To reduce the effect of local minima, we run the majorization algorithm using ten random starts. A computer program that implements the majorization algorithm is available online.[3]

2.6 Step 6: Visualization

The sixth step is the visualization of the low-dimensional space that results from the fifth step. The low-dimensional space should be visualized in such a way that it gives an overview and understanding of the structure of the knowledge domain under consideration.

A straightforward approach to visualize the low-dimensional space is by displaying a label at the computed location of an item. This basic approach is still taken in many studies in the literature on KDViz. Another approach that can be taken is to visualize the low-dimensional space as an undirected graph. In this approach, a label is displayed at the computed location of an item and an edge is generated if the strength between two items is above a certain threshold value.

Figure 1 shows a graph-based visualization of the 252 concepts that were positioned in a two-dimensional space in the previous step. As one can see, the large number of overlapping concept labels and edges obscure the structure of the visualization and make it rather incomprehensible. It is for example difficult or almost impossible to differentiate the individual concepts and to investigate the distances between the concepts. As a consequence, it is not easy to get an overview and understanding of the important topics of

[3] http://www.neesjanvaneck.nl/vos/.

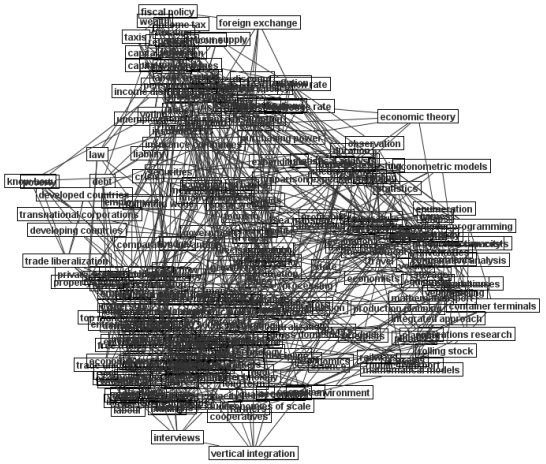

Figure 1. Graph-based visualization of concept associations.

Figure 2. Visualization of concept associations using non-overlapping labels and transparent circles.

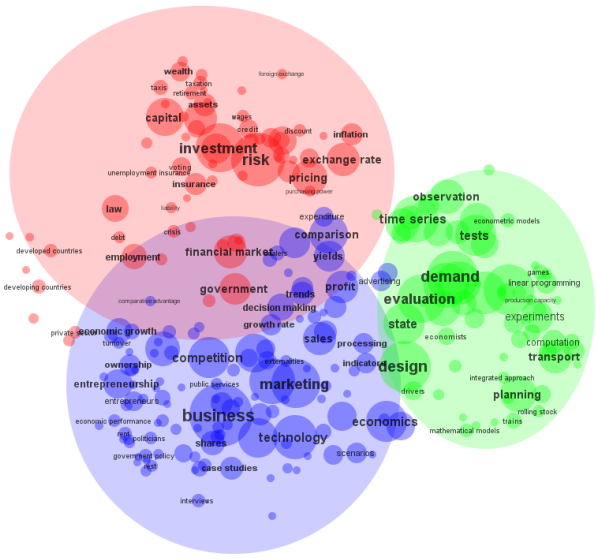

Figure 3. Cluster-based visualization of concept associations.

the *Economics & Management* part of the RePub repository and the relation between these topics. In the next section, we propose a visualization approach that attempts to overcome the above-mentioned problems.

3 Cluster-Based Visualization

As demonstrated in the previous section, straightforward visualization approaches that fit in the KDViz process may produce incomprehensible visualizations with many overlapping labels and connections. As a result, the visualizations do not provide a good overview and understanding of the structure of the knowledge domain under consideration.

Our first attempt to make the visualization in Figure 1 more comprehensible dealt with the presentation of the 252 concepts. First of all we prevented that concept labels overlap each other in the visualization. This was done by displaying only the labels of the most relevant concepts and by displaying these labels only if they are non-overlapping. We computed the relevance of the concepts based on their occurrence frequency in the corpus of abstracts. As a consequence, the labels of some concepts are not displayed. It is nevertheless interesting to display the location of these concepts, since the distribution of all concept locations provides insight into the structure of the visualization. Therefore we displayed the location of a concept using some shape. After experimenting with different shapes we decided to use circles instead of rectangles, as circles reduced the visual cluttering. In order not to put too much attention on just a single concept we made the circles transparent. In order to visually emphasize relevance we used bigger circles for more relevant concepts and smaller circles for less relevant concepts.

Figure 2 shows the resulting visualization. First of all the visualization looks less chaotic than the visualization presented in Figure 1. Moreover, the visualization provides us with a first overview of the most relevant concepts of the analyzed domain knowledge. To gain more insight into the visualization presented in Figure 2, the amount and complexity of the presented information needs to be reduced further. In order to do this, we investigated the use of concept clustering.

We chose to use hierarchical clustering [7] because one does not have to specify the number of clusters in advance and because it returns a hierarchy of clusters instead of a set of flat clusters, as is the case for most other clustering methods, e.g., k-means. We used bottom-up hierarchical clustering (also called agglomerative clustering) starting from clusters of size one (containing one concept) and further aggregating clusters based on their proximity (second level clustering, third level clustering, etc.). We tried out several methods to compute the distances between clusters: complete-linkage, single-linkage, and average-linkage. Based on the cluster analysis that we subsequently performed we decided to use average linkage as it produced the best results for our data. To represent the different clusters in the visualization, we displayed the circles belonging to concepts from different clusters using different colors. In addition, we showed large transparent ovals in the color of

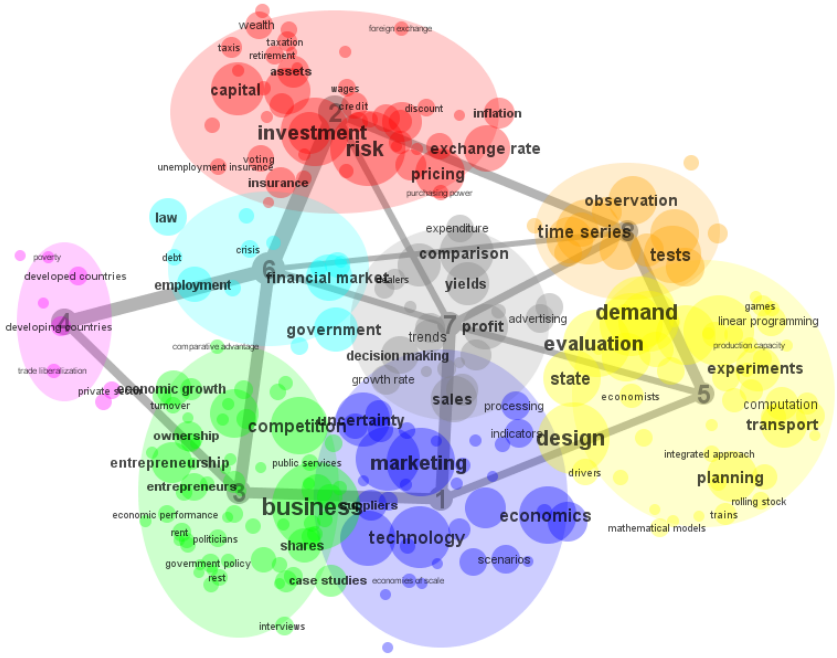

Figure 4. Cluster-based visualization of concept associations including edge-bundles.

Table 2. Most frequently occurring concepts per cluster.

Cluster 1 (●)	Cluster 2 (●)	Cluster 3 (●)	Cluster 4 (●)	Cluster 5 (●)	Cluster 6 (●)	Cluster 7 (●)	Cluster 8 (●)
marketing	risk	business	developing countries	demand	financial market	comparison	time series
interest	investment	competition	developed countries	design	government	sales	tests
technology	capital	trade	private sector	production	law	profit	testing
economics	pricing	entrepreneurship	property rights	evaluation	employment	goods	observation
managers	income	economic growth	trade liberalization	efficiency	banking	yields	methodology
uncertainty	exchange rate	shares	poverty	probability	crisis	decision making	forecasts
discipline	unemployment	case studies	know how	supply	budget	trends	statistics
science	interest rate	productivity	transnational corporations	state	minimum wage	promotion	density
suppliers	assets	ownership		planning	debt	weight	sampling
indicators	insurance	entrepreneurs		measurement	new products	growth rate	duration

a cluster around most of the cluster's concepts in order to better emphasize the clustering.

Figure 3 shows a cluster-based visualization containing three clusters. We believe that the visualization gives an accurate high-level overview of the analyzed domain knowledge. As mentioned earlier, the analyzed domain knowledge consists of abstracts of publications taken from the *Economics & Management* section of the RePub repository. The authors of these publications are affiliated with the Erasmus University Rotterdam and, more importantly, most of them are active in the fields of economics, management and business, or operations research. By inspecting the concepts in each of the three clusters in Figure 3 we concluded that, more or less, the red cluster contains concepts from the field of economics, the blue cluster contains concepts from the field of management and business, and the green cluster contains concepts from the field of operations research.

In order to reduce the visual cluttering due to edge drawing, as we saw in Figure 1, the edges between concepts belonging to two separate clusters are bundled in a single edge which connects the centers of the corresponding clusters. The thickness of an edge between two clusters is proportional to its strength, which is calculated as the average association strength of all pairs of concepts from the two connected clusters. In order not to disturb the visualization too much, only those edges which have a strength above a certain threshold value are displayed.

Figure 4 shows a cluster-based visualization containing eight clusters that includes edges between the clusters. Each

of the eight clusters has been numbered in the figure. The concepts in each of the eight clusters give an indication of the topics of the clusters. Table 2 lists the most frequently occurring concepts per cluster. By inspecting and analyzing the visualization in Figure 4 in detail we were able to find out that the discovered clusters correspond in a natural way to research topics in the fields of economics, management and business, and operations research. We labeled the clusters as follows: (1, ●) marketing, (2, ●) investment and risk, (3, ●) business (with emphasis on entrepreneurship), (4, ●) macroeconomics (with emphasis on development countries), (5, ●) operations research, (6, ●) finance, (7, ●) microeconomics, and (8, ●) econometrics. It should be noted that for some clusters the topic was somewhat ambiguous and that therefore the labels may not completely cover the contents of the clusters. By considering the proximity of clusters and the edges between clusters in Figure 4, some additional observations can be made. First, it can be seen that the topic of marketing is located close to the topic of business. This observation makes sense, since marketing and business are two closely related research fields. The observation is also in agreement with the visualization in Figure 3. In this figure the topics of marketing (●) and business (●) both belong to the blue cluster, which represents the field of management and business, as we discussed earlier. Likewise, in Figure 4 it can be seen that the topic of finance (●) is located close to the topic of investment and risk (●). Again, this is in agreement both with what one would naturally expect and with the visualization in Figure 3. In this figure the two topics both belong to the red cluster, which represents the field of economics. More similar observations can be made, but due to space constraints we cannot discuss them all in detail.

4 Conclusions

Two techniques that can help understanding the visualization of concepts spaces in a scientific domain are concept clustering and edge bundling. Our approach uses a hierarchical bottom-up algorithm for clustering because of the advantage that one does not need to specify a priori the number of clusters. Also the chosen clustering method enables us to understand the hierarchical structure of our domain. For edge bundling we aggregate edges between two clusters as a single edge connecting the clusters centers. The thickness of the aggregated edge depends on average association strength of the concepts in the two connected clusters.

We experimented with our approach by visualizing the associations between concepts extracted from a corpus containing abstracts in the field of economics and management. Using concept clustering and edge-bundling we were able to produce comprehensible visualizations that provide a quick overview and understanding of the (hierarchical) structure of the knowledge contained in the corpus.

In the future we would like to provide zooming facilities, which would allow us to navigate through the clustering hierarchy and also inspect the edges involved in an edge bundle. We further plan to enhance the visualization of individual clusters by displaying them as densities [8] rather than as transparent ovals. We also plan to investigate advanced edge bundling techniques, such as hierarchical edge bundling [1, 6], by exploiting the concept hierarchy given in the used concept taxonomy. Finally, we consider further evaluating our approach by applying it to other corpora from the same domain and by comparing the obtained results with the ones presented in this paper.

References

[1] M. Balzer and O. Deussen. Level-of-detail visualization of clustered graph layouts. In *Proc. of the Asia-Pacific Symposium on Visualisation*, pages 133–140. IEEE Computer Society, 2007.

[2] I. Borg and P. J. F. Groenen. *Modern Multidimensional Scaling*. Springer, 2nd edition, 2005.

[3] K. Börner, C. Chen, and K. W. Boyack. Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37:179–255, 2003.

[4] C. Chen. *Mapping Scientific Frontiers: The Quest for Knowledge Visualization*. Springer, 2003.

[5] C. Chen. *Information Visualisation: Beyond the Horizon*. Springer, 2nd edition, 2006.

[6] D. H. R. Holten. Hierarchical edge bundles: visualization of adjacency relations in hierarchical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):741–748, 2006.

[7] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[8] N. J. van Eck, F. Frasincar, and J. van den Berg. Visualizing concept associations using concept density maps. In *Proc. of the 10th International Conference on Information Visualisation*, pages 270–275. IEEE Computer Society, 2006.

[9] N. J. van Eck and L. Waltman. Bibliometric mapping of the computational intelligence field. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 15(5):625–645, 2007.

[10] N. J. van Eck and L. Waltman. VOS: a new method for visualizing similarities between objects. In *Advances in Data Analysis: Proc. of the 30th Annual Conference of the German Classification Society*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 299–306. Springer, 2007.

[11] N. J. van Eck, L. Waltman, and J. van den Berg. A novel algorithm for visualizing concept associations. In *Proc. of the 16th International Workshop on Database and Expert Systems Applications*, pages 405–409. IEEE Computer Society, 2005.