# Induction of the common-sense hierarchies in lexical data

Julian Szymański[1] and Włodzisław Duch[2,3]

[1] Department of Computer Systems Architecture, Gdańsk University of Technology, Poland,
julian.szymanski@eti.pg.gda.pl
[2] Department of Informatics, Nicolaus Copernicus University, Toruń, Poland
[3] School of Computer Engineering, Nanyang Technological University, Singapore
Google: W. Duch

**Abstract.** Unsupervised organization of a set of lexical concepts that captures common-sense knowledge inducting meaningful partitioning of data is described. Projection of data on principal components allow for identification of clusters with wide margins, and the procedure is recursively repeated within each cluster. Application of this idea to a simple dataset describing animals created hierarchical partitioning with each clusters related to a set of features that have common-sense interpretation.

## 1 Introduction

Categorization of concepts into meaningful hierarchies lies at the foundation of understanding their meaning. Ontologies provide such hand-crafted hierarchical classification, but they are based usually on expert knowledge, not on the common-sense knowledge. For example, most biological taxonomies are hard to understand for lay people. There is no relationship between linguistic labels and their referents, so words may only point at the concept, inducing brain states that contain semantic information, predisposing people to meaningful associations and answers. In particular visual similarity is not related to names. Dog's breeds are categorized depending on their function, like Sheepdogs and Cattle Dogs, Scenthounds, Pointing Dogs, Retrievers, Companion and Toy Dogs, with many diverse breeds within each category. Such categories may have very little in common when all properties are considered. Differences between two similar dog breeds may be based on rare traits, not relevant to general classification. This makes identification of objects by their description quite difficult and the problem of forming common sense natural categories worth studying.

In this paper we have focused on relatively simple data describing animals. First, this is a domain where everyone has relatively good understanding of similarity and hierarchical description, second there is a lot of structured information in the Internet resources that may be used to create detailed description of the animals, third one can test the quality of such data by playing word games. We shall look at the novel way of using principal component analysis (PCA) to create hierarchical descriptions, but many other choices and other knowledge domains (for example, automatic classification of library subjects) may be treated using similar methodology.

## 2 The data

The data used in the experiments has been obtained using automatic knowledge acquisition followed by corrections resulting from the 20-questions word game [1]. The point of this game is to guess the concept the opponent is thinking of by asking questions that should narrow down the set of likely concepts. In our implementation[1] the program is trying to make a guess asking people questions. Results are used to correct lexical knowledge and in its final stage controlled dialog between human and computer, based on several plausible scenarios, is added to acquire additional knowledge. If the program wins, guessing the concept correctly, it will strengthen the knowledge related to this concept. If it fails, human is asked additional question „What did you think of?" and concepts related to the answer are added or features are modified according to the information harvested during the game.

Implementation of our knowledge acquisition system based on the 20-question game uses a semantic memory model [2] to store lexical concepts. This approach makes it more versatile than using just correlation matrix, as it has been successfully done in the implementation of this word game[2]. The matrix stores correlations between objects and features using weights that describe mutual association derived from thousands of games, providing decomposition of each concepts into a sum of contributions from questions. Such representation is flat and does not treat lexical features as natural language concepts that allow for creation of a hierarchy of the common sense objects. Our program, based on semantic memory representation, shows elementary linguistic competence collecting common sense knowledge in restricted domains [1], and the knowledge generated may be used in many ways, for example by generating word puzzles.

This lexical data in semantic memory may be reorganized in a way that will introduce generalizations and increase cognitive economy [3]. This hierarchy is induced searching for the directions with highest variance using PCA eigenvectors, separating subsets of concepts and repeating the process to create consecutive subspaces. To illustrate and better understand this process a relatively small experiment has been performed.

A test dataset with 84 concepts (animals, or in general some objects) described by 71 features has been constructed after performing 346 games. The dataset used in the experiments is displayed using Self-Organizing Map (SOM) [4] visualization in Fig. 1 and with parametric Multidimensional Scaling (MDS) [5] in Fig. 2. Distances between points that represent dissimilarities between animals are calculated using cosine measures $d(\mathbf{X}, \mathbf{Z}) = \mathbf{X} \cdot \mathbf{Z} / ||\mathbf{X}|| ||\mathbf{Z}||$

## 3 PCA directions

Expert taxonomies are frequently based on single feature, such as mammals, and then marsupials, but common-sense categorization is based on combination of features that makes objects similar. Principal Component Analysis [6] finds directions of highest data variance. Projecting the data on these direction shows interesting combination of

[1] http://diodor.eti.pg.gda.pl
[2] http://www.20-q.net

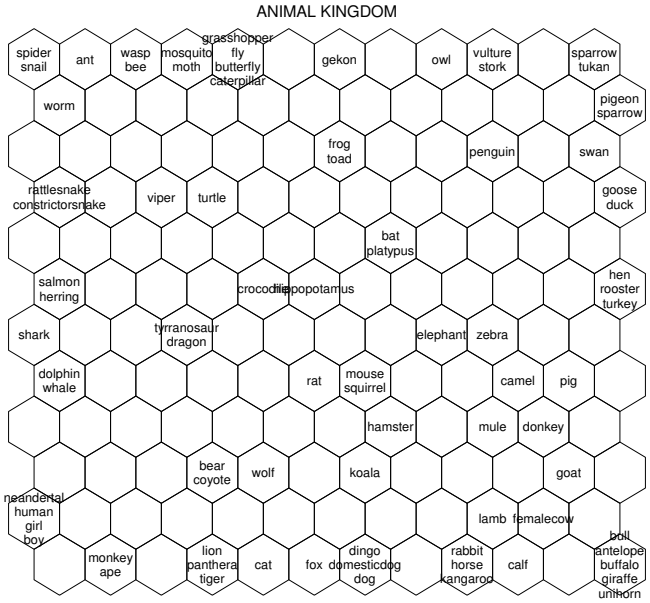


**Fig. 1.** Data used in the experiments visualized with Self-Organizing Map

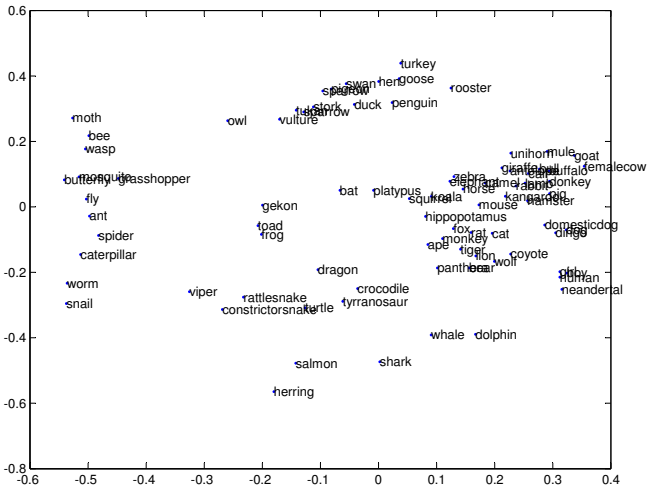


**Fig. 2.** The data used in the experiments visualized with MDS

features and thus helps to select groups of correlated features that separate data points, creating subsets of animals.

A pair of PCA directions may be used for visualization of the data. Projection on the first two directions with largest variance is shown in Fig. 3. The three visualizations (Figure 1, 2, 3) show different aspects of the data. Note for example the cluster

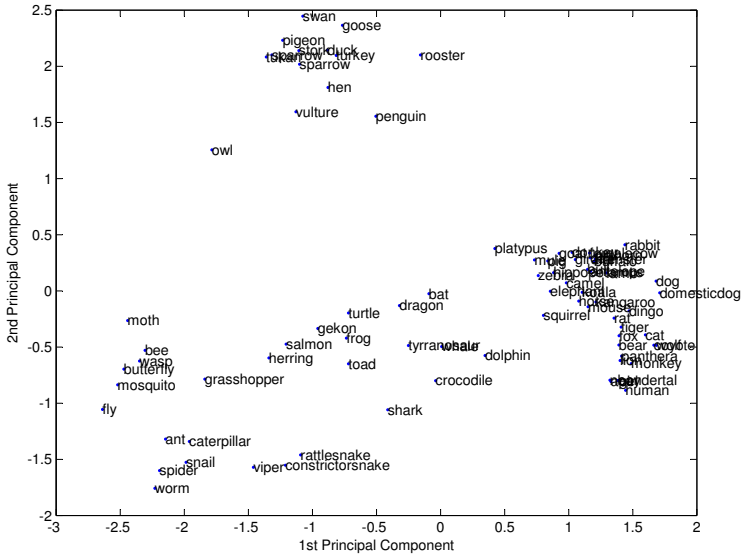


**Fig. 3.** Dataset visualization using two highest Principal Components

formed in SOM containing *lion, pantera* and *tiger*. MDS shows their similarity but can still distinguish between them, while in PCA projection some objects appear between them (*fox, bear*) that do not fall into that cluster. PCA is able to find groups of related features and thus extract some commonsense knowledge approximating meaningful directions in the feature space. At the one end of the axis objects that have a mixture of features making them similar to each other are placed, on the other end objects that do not have such features. Fig. 4 shows coefficients of features in our semantic space for the first six principal components. Each feature, such as *lay-eggs*, *is-mammal* is placed above the line in one of the 6 columns, one for each component, to indicate the value of its coefficient in PCA vector. The most important features (having the highest absolute coefficient weights) in terms of data partitioning can be obtained from subsequent components. In the first vector (lowest row) the most negative (leftmost) coefficients correspond to features *lay-eggs*, *has-wings* describing insects and birds, while the most positive (right-most) are for *is-mammal*, *has-teeth*, *has-coat*, *is-warmblooded*, and others typical for mammals. The second PCA component has most positive coefficients for *has-beak, has-bill, has-feathers, is-bird, has-wings* indicating that this group of features is characteristic for the birds.

Hierarchical clusterization for such groups of features should show interesting commonsense clusters. In Fig. 5 direct projection of all vectors describing animals on each of these principal directions is shown. These projections show different aspects of the data, for example the projection on the second PCA shows a clear cluster for birds, starting with *swan* and ending with *owl* as less typical bird, the third cluster starts with *vulture* and groups other hunting animals. Projection on each PCA component may be used to generate different partitions of all objects.

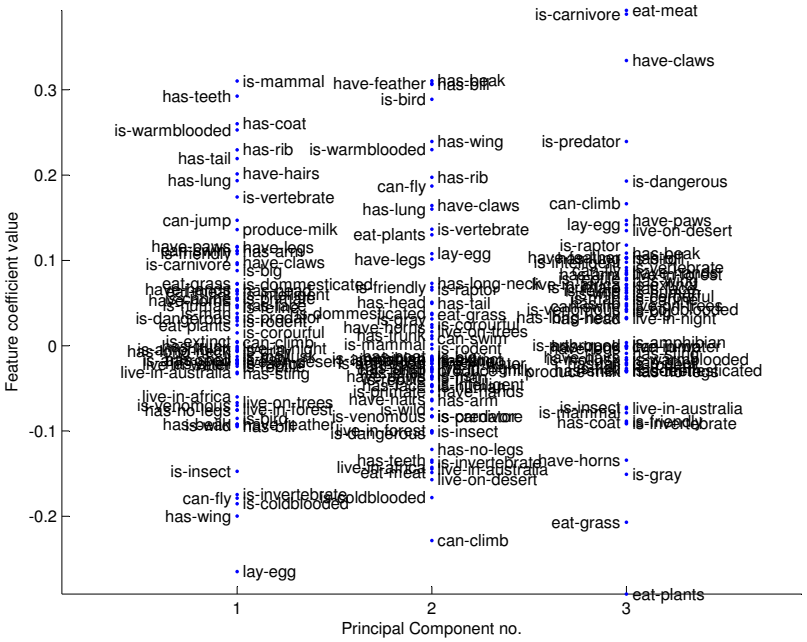


**Fig. 4.** Groups of features related to the principal components

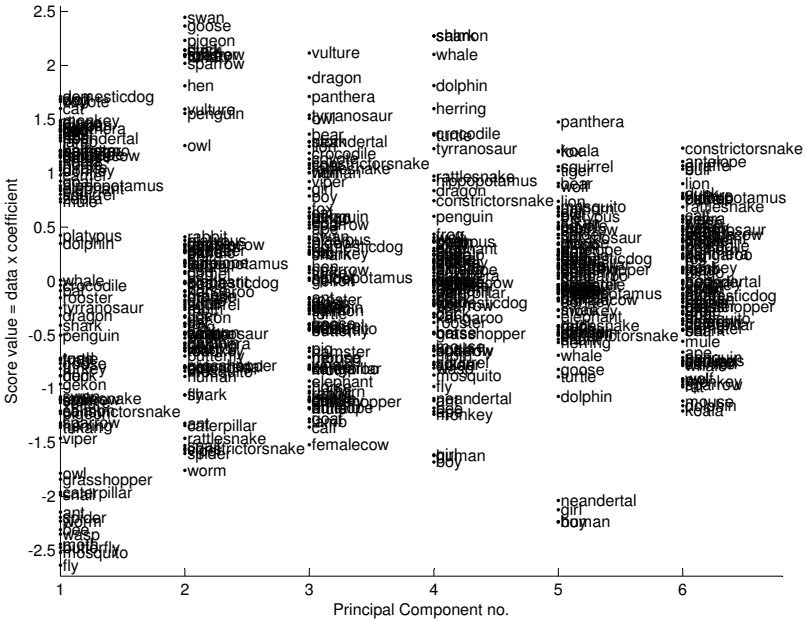


**Fig. 5.** Projections of the data on the first 6 principal components

# 4 Creating hierarchical partitioning

## 4.1 Hierarchical agglomerative partitioning

Creating a hierarchy based on similarity is one of the most effective ways for presenting large sets of concepts. Clustering data using agglomerative approach [7] is most frequently used for showing hierarchical organization of the data. The bottom-up approach using average linkage between clusters on each hierarchy level is shown in Fig. 6.

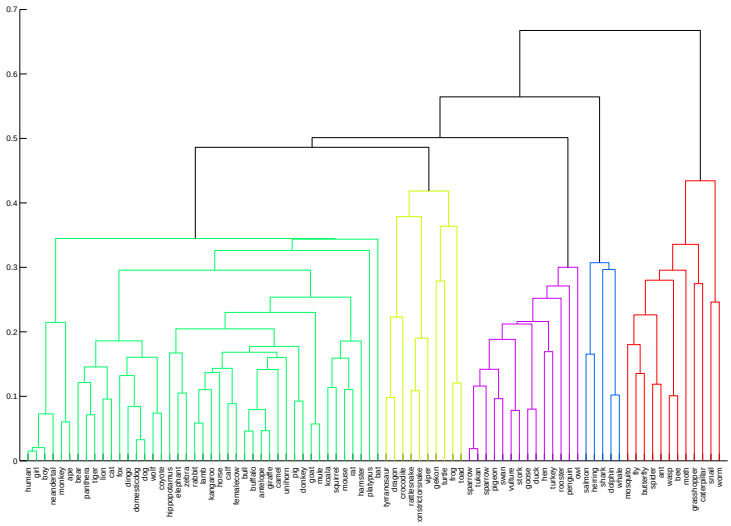


**Fig. 6.** Dendogram for animal kingdom dataset

Hierarchical agglomerative clustering using bottom-up approach binds together groups of objects in a way that frequently does not agree with intuitive partitioning. Moreover, the features used to construct a cluster are not easily traceable.

## 4.2 Hierarchical partitioning with principal components

The distribution of the data points using the first 6 principal components (Fig. 5 ) shows a large gap between two groups projected with the second principal component. These two clusters of data are separated with the largest margin and thus should be meaningful. Hierarchical organization of the data can be analyzed from the point of view of graph theory. In terms of the graph bisection the second eigenvector is most important [8], allowing for creation of normalized cut (partition of the vertices of a graph into two disjoint subsets) [9]. Thus selecting the second principal component is a good start

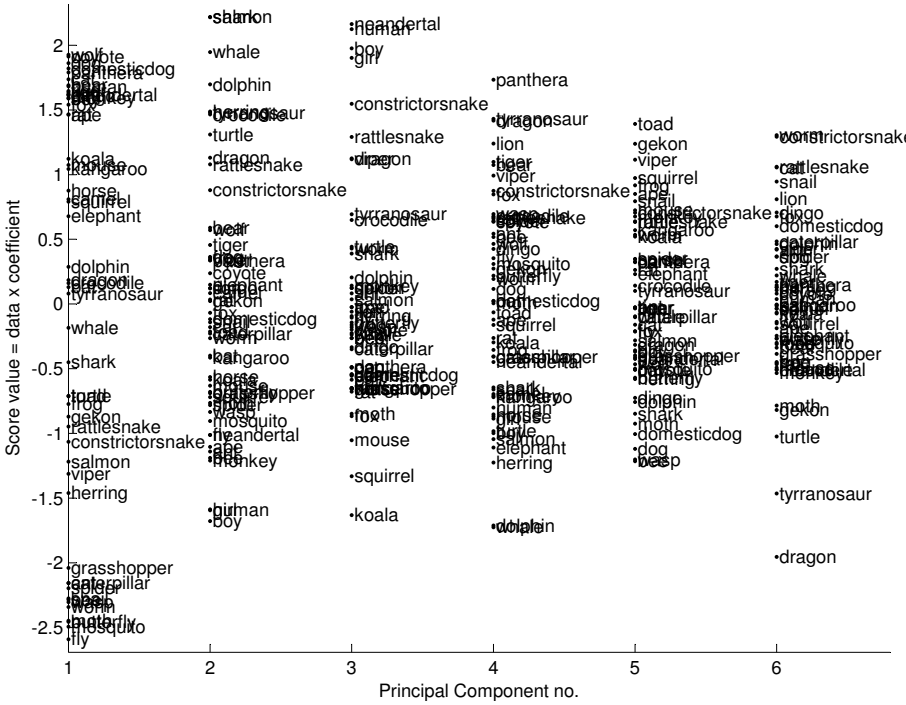


**Fig. 7.** Projections of the reduced data set using succeeding single principal components

to construct hierarchical partitioning. The typical approaches to spectral clustering employs second (biggest) component [10] (that minimize graph conductance) or a second smallest component [11] (due to Rayleigh theorem).

Analyzing subsequent PCA component projections (given in Fig. 5 and 7) shows that the second principal component does not lead always to the best cut in the graph. It is better to select the component that produces the widest separation margin within the data, choosing a different principal component for each hierarchy level. For creating the first hierarchy level the second component is selected, separating birds from other animals, creating one pure and one mixed cluster (Fig. 5). Features of the second PCA component (Fig. 4) with lowest and highest weights include: (-)*climb,* (-)*cold-blooded* and (+)*beak*, (+)*feather*, (+)*bird*, (+)*wing*, (+)*warmblooded*. Note that one feature *is-bird* alone is sufficient to create this partitioning but correlated features separate this cluster in a better way.

To capture some common-sense knowledge hierarchical partitioning is created in a top-down way Each of the newly created clusters is analyzed using PCA and principal components that give the widest separation margins are selected for data partitioning. PCA is performed recursively on reduced data that belong to the selected cluster. In Fig. 7 the first 6 components computed for the large mixed cluster (that does not contain birds) created on the second hierarchy level is presented. This cluster has been formed after separating the birds and other animals with the second component (shown in Fig-


**Fig. 8.** Hierarchy of the data and features used to create it

ure 5). Within this cluster the widest margin is created with the first component and it separate mammals (with the exception of dolphins and whales) from other animals.

Repeating the process described above hierarchical organization of the data is introduced. In Fig. 8 a top part of created hierarchy is shown. At each level of the hierarchy most important features used to create this partition are also displayed.

## 5 Discussion and future directions

An approach to create hierarchical commonsense partitioning of data using recursive Principal Component Analysis has been presented. Results of this procedure have been illustrated on simple data describing animals created using the 20-questions game that is based on model of semantic memory [1]. This approach has been used for creating general clusters within the semantic memory model that stores natural language concepts. Such analysis allows for finding additional correlations between features facilitating associative processes for existing concepts, and improving the learning proces when new information is added to the system. In the neurolinguistic approach to the natural language processing [12] it has been conjectured that the right brain hemisphere

creates receptive fields (called “cosets", or constraint-sets) that constrain semantic interpretation, although they do not have linguistic labels themselves. The process describe here may be an approximation ot some of the neural processes responsible for language comprehension.

Hierarchical organization of lexical data has been created here in an unsupervised way by selecting linear combinations of features that provide clear separation of concepts. Extension of this approach may be based on bi-clustering, taking into account clusters of features that are relevant for creating meaningful clusters of data. The main idea is to strengthen features that are correlated to the dominant one, or to the features given by the user who may want to view the data from a specific angle [13]. Non-negative matrix factorization [14] is another useful technique that may replace PCA. Many other variants of unsupervised data analysis methods are worth exploring in the context of this approach to induction of the common-sense hierarchies in data.

## ACKNOWLEDGEMENTS

The work has been supported by the Polish Ministry of Science and Higher Education under research grant N519 432 338.

## References

1. Szymański, J., Duch, W.: Information retrieval with semantic memory model. Cognitive Systems Research **(in print)** (2011)
2. Tulving, E.: Episodic and semantic memory. Organization of memory (1972) 381–402
3. Conrad, C.: Cognitive economy in semantic memory. (1972)
4. Kohonen, T.: The self-organizing map. Proceedings of the IEEE **78** (1990) 1464–1480
5. Shepard, R.: Multidimensional scaling, tree-fitting, and clustering. Science **210** (1980) 390
6. Jolliffe, I.: Principal component analysis. Wiley Online Library (2002)
7. Day, W., Edelsbrunner, H.: Efficient algorithms for agglomerative hierarchical clustering methods. Journal of classification **1** (1984) 7–24
8. Rahimi, A., Recht, B.: Clustering with normalized cuts is clustering with a hyperplane. Statistical Learning in Computer Vision (2004)
9. Dhillon, I.: Co-clustering documents and words using bipartite spectral graph partitioning. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2001) 269–274
10. Kannan, R., Vetta, A.: On clusterings: Good, bad and spectral. Journal of the ACM (JACM) **51** (2004) 497–515
11. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence **22** (2000) 888–905
12. Duch, W., Matykiewicz, P., Pestian, J.: Neurolinguistic approach to natural language processing with applications to medical text analysis. Neural Networks **21(10)** (2008) 1500–1510
13. Szymański, J., Duch, W.: Dynamic Semantic Visual Information Management. Proceedings of the 9th International Conference on Information and Management Sciences (2010) 107–117
14. D.D., L., S., S.: Learning the parts of objects by non-negative matrix factorization. Nature **401** (1999) 788–791