Vortrag gehalten bei: 51st Annual Meeting of the Association for Computational Linguistics, 4-9 August, Sofia, Bulgaria / Association for Computational Linguistics

A Visual Analytics System for Cluster Exploration

Andreas Lamprecht[1], **Annette Hautli**[2], **Christian Rohrdantz**[1], **Tina Bögel**[2]
[1]Department of Computer Science, [2]Department of Linguistics
University of Konstanz, Germany
{firstname.lastname}@uni-konstanz.de

Abstract

This paper offers a new way of representing the results of automatic clustering algorithms by employing a Visual Analytics system which maps members of a cluster and their distance to each other onto a two-dimensional space. A case study on Urdu complex predicates shows that the system allows for an appropriate investigation of linguistically motivated data.

1 Motivation

In recent years, Visual Analytics systems have increasingly been used for the investigation of linguistic phenomena in a number of different areas, starting from literary analysis (Keim and Oelke, 2007) to the cross-linguistic comparison of language features (Mayer et al., 2010a; Mayer et al., 2010b; Rohrdantz et al., 2012a) and lexical semantic change (Rohrdantz et al., 2011; Heylen et al., 2012; Rohrdantz et al., 2012b). Visualization has also found its way into the field of computational linguistics by providing insights into methods such as machine translation (Collins et al., 2007; Albrecht et al., 2009) or discourse parsing (Zhao et al., 2012).

One issue in computational linguistics is the interpretability of results coming from machine learning algorithms and the lack of insight they offer on the underlying data. This drawback often prevents theoretical linguists, who work with computational models and need to see patterns on large data sets, from drawing detailed conclusions. The present paper shows that a Visual Analytics system facilitates "analytical reasoning [...] by an *interactive* visual interface" (Thomas and Cook, 2006) and helps resolving this issue by offering a customizable, in-depth view on the statistically generated result and simultaneously an at-a-glance overview of the overall data set.

In particular, we focus on the visual representation of automatically generated clusters, in itself not a novel idea as it has been applied in other fields like the financial sector, biology or geography (Schreck et al., 2009). But as far as the literature is concerned, *interactive* systems are still less common, particularly in computational linguistics, and they have not been designed for the specific needs of theoretical linguists. This paper offers a method of visually encoding clusters and their internal coherence with an interactive user interface, which allows users to adjust underlying parameters and their views on the data depending on the particular research question. By this, we partly open up the "black box" of machine learning.

The linguistic phenomenon under investigation, for which the system has originally been designed, is the varied behavior of nouns in N+V CP complex predicates in Urdu (e.g., memory+do = 'to remember') (Mohanan, 1994; Ahmed and Butt, 2011), where, depending on the lexical semantics of the noun, a set of different light verbs is chosen to form a complex predicate. The aim is an automatic detection of the different groups of nouns, based on their light verb distribution. Butt et al. (2012) present a static visualization for the phenomenon, whereas the present paper proposes an interactive system which alleviates some of the previous issues with respect to noise detection, filtering, data interaction and cluster coherence. For this, we proceed as follows: section 2 explains the proposed Visual Analytics system, followed by the linguistic case study in section 3. Section 4 concludes the paper.

2 The system

The system requires a plain text file as input, where each line corresponds to one data object.In our case, each line corresponds to one Urdu noun (data object) and contains its unique ID (the name of the noun) and its bigram frequencies with the

four light verbs under investigation, namely *kar* 'do', *ho* 'be', *hu* 'become' and *rakH* 'put'; an exemplary input file is shown in Figure 1.

From a data analysis perspective, we have four-dimensional data objects, where each dimension corresponds to a bigram frequency previously extracted from a corpus. Note that more than four dimensions can be loaded and analyzed, but for the sake of simplicity we focus on the four-dimensional Urdu example for the remainder of this paper. Moreover, it is possible to load files containing absolute bigram frequencies and relative frequencies. When loading absolute frequencies, the program will automatically calculate the relative frequencies as they are the input for the clustering. The absolute frequencies, however, are still available and can be used for further processing (e.g. filtering).

file with relative frequencies

```
ID,kar,ho,hu,rakH
kAm,0.953,0.047,0.000,0.000
h2As3il,0.771,0.222,0.007,0.000
*a2*lAn,0.982,0.011,0.007,0.000
bAt,0.853,0.147,0.000,0.000
SurUa2,0.530,0.384,0.086,0.000
*s*t*a2*mAl,0.873,0.121,0.006,0.000
p<ye>S,0.864,0.131,0.005,0.000
```

file with absolute frequencies

```
ID,kar,ho,hu,rakH
دس,2,0,0,0
مان,37,66,2,7
نيس,10,0,0,0
تيبرت,16,3,0,0
اتكساج.,1,0,0,0
يڤصيف,119,20,0,0
متب,1,1,0,0
```

Figure 1: preview of appropriate file structures

2.1 Initial opening and processing of a file

It is necessary to define a metric distance function between data objects for both clustering and visualization. Thus, each data object is represented through a high dimensional (in our example four-dimensional) numerical vector and we use the Euclidean distance to calculate the distances between pairs of data objects. The smaller the distance between two data objects, the more similar they are.

For visualization, the high dimensional data is projected onto the two-dimensional space of a computer screen using a principal component analysis (PCA) algorithm[1]. In the 2D projection, the distances between data objects in the high-dimensional space, i.e. the dissimilarities of the bigram distributions, are preserved as accurately as possible. However, when projecting a high-dimensional data space onto a lower dimension, some distinctions necessarily level out: two data objects may be far apart in the high-dimensional space, but end up closely together in the 2D projection. It is important to bear in mind that the 2D visualization is often quite insightful, but interpretations have to be verified by interactively investigating the data.

The initial clusters are calculated (in the high-dimensional data space) using a default k-Means algorithm[2] with k being a user-defined parameter. There is also the option of selecting another clustering algorithm, called the Greedy Variance Minimization[3] (GVM), and an extension to include further algorithms is under development.

2.2 Configuration & Interaction

2.2.1 The main window

The main window in Figure 2 consists of three areas, namely the configuration area (a), the visualization area (b) and the description area (c). The visualization area is mainly built with the piccolo2d library[4] and initially shows data objects as colored circles with a variable diameter, where color indicates cluster membership (four clusters in this example). Hovering over a dot displays information on the particular noun, the cluster membership and the light verb distribution in the description area to the right. By using the mouse wheel, the user can zoom in and out of the visualization.

A very important feature for the task at hand is the possibility to select multiple data objects for further processing or for filtering, with a list of selected data objects shown in the description area. By right-clicking on these data objects, the user can assign a unique class (and class color) to them. Different clustering methods can be employed using the options item in the menu bar.

Another feature of the system is that the user can fade in the cluster centroids (illustrated by a larger dot in the respective cluster color in Figure 2), where the overall feature distribution of the cluster can be examined in a tooltip hovering over the corresponding centroid.

2.2.2 Visually representing data objects

To gain further insight into the data distribution based on the 2D projection, the user can choose between several ways to visualize the individual data objects, all of which are shown in Figure 3. The standard visualization type is shown on the left and consists of a **circle** which encodes cluster membership via color.

[1] http://workshop.mkobos.com/2011/java-pca-transformation-library/

[2] http://java-ml.sourceforge.net/api/0.1.7/ (From the JML library)

[3] http://www.tomgibara.com/clustering/fast-spatial/

[4] http://www.piccolo2d.org/

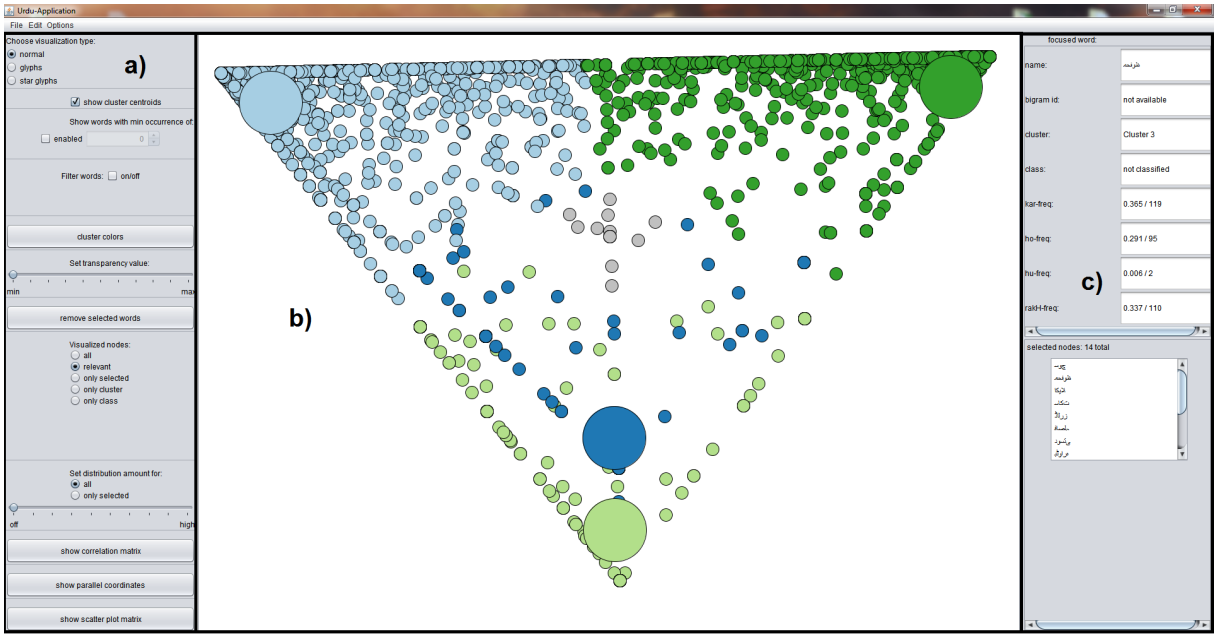


Figure 2: Overview of the main window of the system, including the configuration area (a), the visualization area (b) and the description area (c). Large circles are cluster centroids.

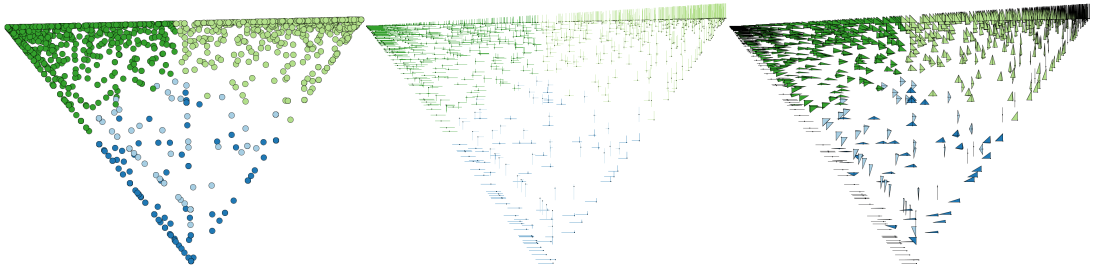
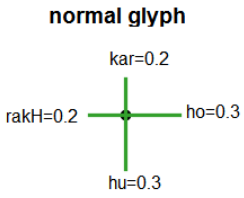
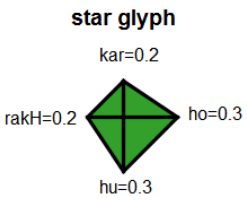


Figure 3: Different visualizations of data points

Alternatively, **normal glyphs** and **star glyphs** can be displayed. The middle part of Figure 3 shows the data displayed with **normal glyphs**. In this view, the relative frequency of each light verb is mapped onto the length of a line. The lines start in north position and are positioned clockwise around the center according to their occurrence in the input file. This view has the advantage that overall feature dominance in a cluster can be seen at-a-glance.





The visualization type on the right in Figure 3 is called the **star glyph**, an extension to normal glyphs. Here, the line endings are connected, forming a "star". As in the representation with the glyphs, this makes similar data objects easily recognizable and comparable with each other.

2.2.3 Filtering options

Our systems offers options for filtering data according to different criteria.

Filter by means of bigram occurrence By activating the bigram occurrence filtering, it is possible to only show those nouns, which occur in bigrams with a certain selected subset of all features (light verbs) only. This is especially useful when examining possible commonalities.

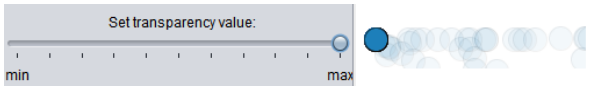
Filter selected words Another opportunity of showing only items of interest is to select and display them separately. The PCA is recalculated for these data objects and the visualization is stretched to the whole area.

Filter selected cluster Additionally, the user can visualize a specific cluster of interest. Again, the PCA is recalculated and the visualization stretched to the whole area. The cluster can then be manually fine-tuned and cleaned, for instance by removing wrongly assigned items.

2.2.4 Options to handle overplotting

Due to the nature of the data, much overplotting occurs. For example, there are many words, which only occur with one light verb. The PCA assigns the same position to these words and, as a consequence, only the top bigram can be viewed in the visualization. In order to improve visual access to overplotted data objects, several methods that allow for a more differentiated view of the data have been included and are described in the following paragraphs.

Change transparency of data objects By modifying the transparency with the given slider, areas with a dense data population can be readily identified, as shown in the following example:

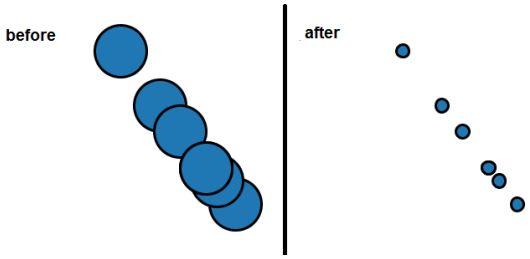


Repositioning of data objects To reduce the overplotting in densely populated areas, data objects can be repositioned randomly having a fixed deviation from their initial position. The degree of deviation can be interactively determined by the user employing the corresponding slider:

The user has the option to reposition either all data objects or only those that are selected in advance.

Frequency filtering If the initial data contains absolute bigram frequencies, the user can filter the visualized words by frequency. For example, many nouns occur only once and therefore have an observed probability of 100% for co-occurring with one of the light verbs. In most cases it is useful to filter such data out.

Scaling data objects If the user zooms beyond the maximum zoom factor, the data objects are scaled down. This is especially useful, if data objects are only partly covered by many other objects. In this case, they become fully visible, as shown in the following example:



2.3 Alternative views on the data

In order to enable a holistic analysis it is often valuable to provide the user with different views on the data. Consequently, we have integrated the option to explore the data with further standard visualization methods.

2.3.1 Correlation matrix

The correlation matrix in Figure 4 shows the correlations between features, which are visualized by circles using the following encoding: The size of a circle represents the correlation strength and the color indicates whether the corresponding features are negatively (white) or positively (black) correlated.

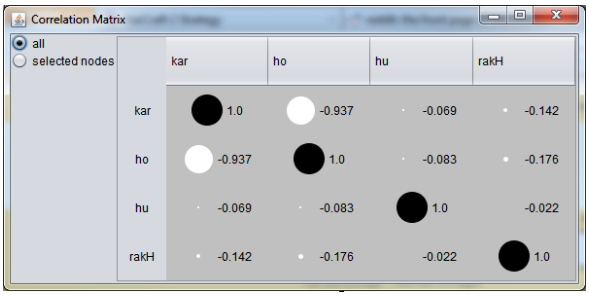


Figure 4: example of a correlation matrix

2.3.2 Parallel coordinates

The parallel coordinates diagram shows the distribution of the bigram frequencies over the different dimensions (Figure 5). Every noun is represented with a line, and shows, when hovered over, a tooltip with the most important information. To filter the visualized words, the user has the option of displaying previously selected data objects, or s/he can restrict the value range for a feature and show only the items which lie within this range.

2.3.3 Scatter plot matrix

To further examine the relation between pairs of features, a scatter plot matrix can be used (Figure 6). The individual scatter plots give further insight into the correlation details of pairs of features.

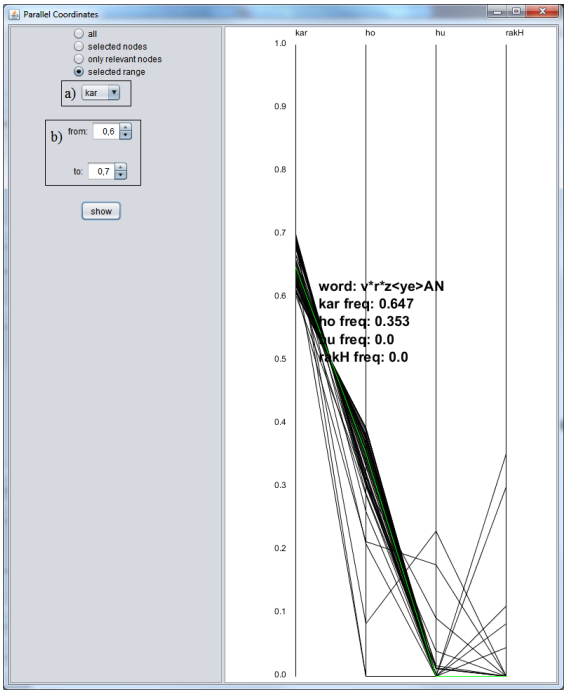


Figure 5: Parallel coordinates diagram

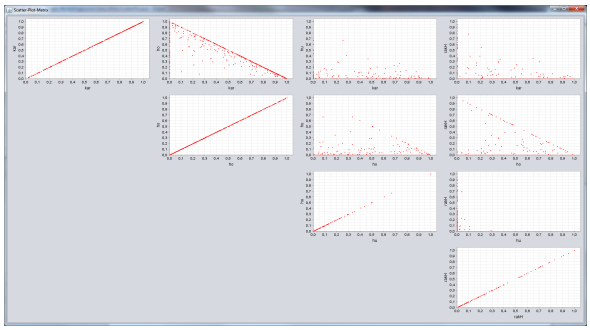


Figure 6: Example showing a scatter plot matrix.

3 Case study

In principle, the Visual Analytics system presented above can be used for any kind of cluster visualization, but the built-in options and add-ons are particularly designed for the type of work that linguists tend to be interested in: on the one hand, the user wants to get a quick overview of the overall patterns in the phenomenon, but on the same time, the system needs to allow for an in-depth data inspection. Both is given in the system: The overall cluster result shown in Figure 2 depicts the coherence of clusters and therefore the overall pattern of the data set. The different glyph visualizations in Figure 3 illustrate the properties of each cluster. Single data points can be inspected in the description area. The randomization of overplotted data points helps to see concentrated cluster patterns where light verbs behave very similarly in different noun+verb complex predicates.

The biggest advantage of the system lies in the ability for interaction: Figure 7 shows an example of the visualization used in Butt et al. (2012), the input being the same text file as shown in Figure 1. In this system, the relative frequencies of each noun with each light verb is correlated with color saturation — the more saturated the color to the right of the noun, the higher the relative frequency of the light verb occurring with it. The number of the cluster (here, 3) and the respective nouns (e.g. *kAm* 'work') is shown to the left. The user does not get information on the coherence of the cluster, nor does the visualization show prototypical cluster patterns.

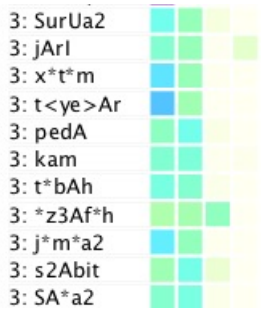


Figure 7: Cluster visualization in Butt et al. (2012)

Moreover, the system in Figure 7 only has a limited set of interaction choices, with the consequence that the user is not able to adjust the underlying data set, e.g. by filtering out noise. However, Butt et al. (2012) report that the Urdu data is indeed very noisy and requires a manual cleaning of the data set before the actual clustering. In the system presented here, the user simply marks conspicuous regions in the visualization panel and removes the respective data points from the original data set. Other filtering mechanisms, e.g. the removal of low frequency items which occur due to data sparsity issues, can be removed from the overall data set by adjusting the parameters.

A linguistically-relevant improvement lies in the display of cluster centroids, in other words the typical noun + light verb distribution of a cluster. This is particularly helpful when the linguist wants to pick out prototypical examples for the cluster in order to stipulate generalizations over the other cluster members.

4 Conclusion

In this paper, we present a novel visual analytics system that helps to automatically analyze bigrams extracted from corpora. The main purpose is to enable a more informed and steered cluster analysis than currently possible with standard methods. This includes rich options for interaction, e.g. display configuration or data manipulation. Initially, the approach was motivated by a concrete research problem, but has much wider applicability as any kind of high-dimensional numerical data objects can be loaded and analyzed. However, the system still requires some basic understanding about the algorithms applied for clustering and projection in order to prevent the user to draw wrong conclusions based on artifacts. Bearing this potential pitfall in mind when performing the analysis, the system enables a much more insightful and informed analysis than standard non-interactive methods.

In the future, we aim to conduct user experiments in order to learn more about how the functionality and usability could be further enhanced.

Acknowledgments

This work was partially funded by the German Research Foundation (DFG) under grant BU 1806/7-1 "Visual Analysis of Language Change and Use Patterns" and the German Federal Ministry of Education and Research (BMBF) under grant 01461246 "VisArgue" under research grant.

References

Tafseer Ahmed and Miriam Butt. 2011. Discovering Semantic Classes for Urdu N-V Complex Predicates. In *Proceedings of the international Conference on Computational Semantics (IWCS 2011)*, pages 305–309.

Joshua Albrecht, Rebecca Hwa, and G. Elisabeta Marai. 2009. The Chinese Room: Visualization and Interaction to Understand and Correct Ambiguous Machine Translation. *Comput. Graph. Forum*, 28(3):1047–1054.

Miriam Butt, Tina Bögel, Annette Hautli, Sebastian Sulger, and Tafseer Ahmed. 2012. Identifying Urdu Complex Predication via Bigram Extraction. In *In Proceedings of COLING 2012, Technical Papers*, pages 409 – 424, Mumbai, India.

Christopher Collins, M. Sheelagh T. Carpendale, and Gerald Penn. 2007. Visualization of Uncertainty in Lattices to Support Decision-Making. In *EuroVis 2007*, pages 51–58. Eurographics Association.

Kris Heylen, Dirk Speelman, and Dirk Geeraerts. 2012. Looking at word meaning. An interactive visualization of Semantic Vector Spaces for Dutch synsets. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 16–24.

Daniel A. Keim and Daniela Oelke. 2007. Literature Fingerprinting: A New Method for Visual Literary Analysis. In *IEEE VAST 2007*, pages 115–122. IEEE.

Thomas Mayer, Christian Rohrdantz, Miriam Butt, Frans Plank, and Daniel A. Keim. 2010a. Visualizing Vowel Harmony. *Linguistic Issues in Language Technology*, 4(Issue 2):1–33, December.

Thomas Mayer, Christian Rohrdantz, Frans Plank, Peter Bak, Miriam Butt, and Daniel A. Keim. 2010b. Consonant Co-Occurrence in Stems across Languages: Automatic Analysis and Visualization of a Phonotactic Constraint. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, pages 70–78, Uppsala, Sweden, July. Association for Computational Linguistics.

Tara Mohanan. 1994. *Argument Structure in Hindi*. Stanford: CSLI Publications.

Christian Rohrdantz, Annette Hautli, Thomas Mayer, Miriam Butt, Frans Plank, and Daniel A. Keim. 2011. Towards Tracking Semantic Change by Visual Analytics. In *ACL 2011 (Short Papers)*, pages 305–310, Portland, Oregon, USA, June. Association for Computational Linguistics.

Christian Rohrdantz, Michael Hund, Thomas Mayer, Bernhard Wälchli, and Daniel A. Keim. 2012a. The World's Languages Explorer: Visual Analysis of Language Features in Genealogical and Areal Contexts. *Computer Graphics Forum*, 31(3):935–944.

Christian Rohrdantz, Andreas Niekler, Annette Hautli, Miriam Butt, and Daniel A. Keim. 2012b. Lexical Semantics and Distribution of Suffixes - A Visual Analysis. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 7–15, April.

Tobias Schreck, Jürgen Bernard, Tatiana von Landesberger, and Jörn Kohlhammer. 2009. Visual cluster analysis of trajectory data with interactive kohonen maps. *Information Visualization*, 8(1):14–29.

James J. Thomas and Kristin A. Cook. 2006. A Visual Analytics Agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13.

Jian Zhao, Fanny Chevalier, Christopher Collins, and Ravin Balakrishnan. 2012. Facilitating Discourse Analysis with Interactive Visualization. *IEEE Trans. Vis. Comput. Graph.*, 18(12):2639–2648.