Visualising latent semantic spaces for sense-making of natural language text

Ana Šemrov[1], Alan F Blackwell[1], Advait Sarkar[1,2]

[1]Computer Laboratory, University of Cambridge, UK
[2]Microsoft Research, Cambridge, UK
ana.semrov@gmail.com, alan.blackwell@cl.cam.ac.uk
advait@microsoft.com

Abstract. Latent Semantic Analysis is widely used for natural language processing, but is difficult to visualise and interpret. We present an interactive visualisation that enables the interpretation of latent semantic spaces. It combines a multi-dimensional scatterplot diagram with a novel clutter-reduction strategy based on hierarchical clustering. A study with 12 non-expert participants showed that our visualisation was significantly more usable than experimental alternatives, and helped users make better sense of the latent space.

1 Introduction

This design case study explores an increasingly common class of diagram, which represents a statistical model used to explore unstructured, qualitative datasets, such as our example dataset: a snapshot of Wikipedia. We focus on Latent Semantic Analysis (LSA), one of a class of methods that represents words as vectors, where dimensions of the vector space capture aspects of word meaning [1,2]. LSA dimensions have been shown to be good predictors of human meaning-based judgements [3], perform well in tasks based on word similarity [4] and are useful in sentiment analysis [5]. Unfortunately, users do not find it easy to interpret the dimensions extracted from LSA.

Our research therefore investigates whether interactive diagrams can be used to provide a more interpretable mapping between a model created using LSA and the domain content of the vocabulary being mapped, and whether a mapping of this kind can provide an effective basis for sensemaking and exploration.

Conventional quantitative graphs are valuable to experts who are interested in understanding and refining the model. It is not unusual for experts in a domain to invent tools that will assist them in their own tasks, and as a result, we find that statistical visualisation approaches are widespread in the data analytics and natural language processing literature. However, such visualisations may be less valuable to those who are not experts.

Our distinctive approach focuses on presenting the semantic relationships between words, treating the problem as one of diagram design. We visualise semantic structure using geometric regions that summarise clusters of related words. The user can explore any word group cluster by selecting and "expanding" the view to focus on those words. Exploration of clusters can be recursive, allowing navigation of a semantic hierarchy. Interaction with lower levels of the hierarchy

allows the user to explore closely related words, while interaction with higher levels provides a thematic overview of the corpus. We use diagrammatic design cues to communicate these different interpretation opportunities to the user.

We demonstrate through a user study that our system improves the ability of non-expert users to discover groups of related words and assign meaning to dimensions, when compared to two more conventional alternative visualisations.

1.1 Related work

Visualisation of multidimensional datasets has been previously explored. ScatterDice uses scatterplots and scatterplot matrices to represent the dataset [6]. An alternative approach uses parallel coordinates and hierarchical clustering [7], lines are coloured according to the proximity of their corresponding data points in a cluster hierarchy. Other approaches to scatterplot matrices, including density contour, sunflower plots, and density estimations, have been compared [8].

A notable prior design, aimed at improving the understanding of latent semantic spaces, is a flattened network visualisation of the space [9]. A separate network can be displayed for each dimension, where the length of edges between words corresponds to the similarity between those words on that dimension.

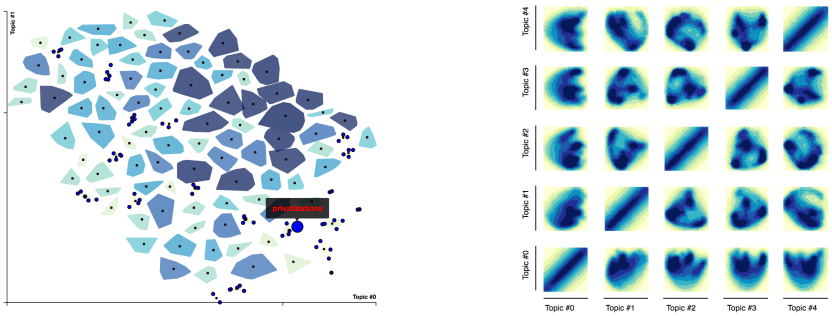
Strategies for clutter-reduction have been well-explored. Some taxonomies distinguish between clutter reduction strategies affecting the appearance of individual data points, those spatially distorting the space to displace the data points, and animation techniques [10]. Another survey presents visual aggregation strategies including multidimensional scatterplots, parallel coordinates, star plots, and a model of hierarchical aggregation related to our approach [11].

2 LSA model construction

The Westbury Lab Wikipedia Corpus [12] was used during development as well as the experiment. This snapshot of the English Wikipedia contains articles as plain text without Wiki markup, links and other non-content material.

After removing stopwords and words occurring in fewer than two documents, we constructed word-document co-occurrence matrix $\mathbf{A}$. Rows correspond to words, columns to documents, and each entry a_{ij} corresponds to the appearances of word i in document j. We applied inverse document frequency (IDF) weighting; words appearing in fewer documents were prioritised relative to common words. We then used a standard LSA library [13]. The co-occurrence matrix is factorised using singular value decomposition [14]. The $n \times m$ matrix $\mathbf{A}$ can be written as the product $\mathbf{A} = U\Sigma V^T$, where U is an orthogonal $n \times n$ matrix that recasts the original row (word) vectors as vectors of n derived orthogonal factors; likewise V is an orthogonal $m \times m$ matrix describing the original columns. Σ is an $n \times m$ diagonal matrix, whose diagonal entries are 'singular values' of the matrix and the columns of U and V are respectively the right and left singular vectors.

The top k singular values, and the corresponding rows and columns from U and V, give a k-rank approximation for A. The word vectors in U can thus be expressed with k dimensions, instead of in terms of every document. The choice of k is task and content dependent [15] and is typically tuned empirically [1]. Using the L-method [16] we found $k = 5$ dimensions sufficient for our corpus.



(a) Multiple clusters visualised. A word is being inspected via tooltip.

(b) Heatmap matrix. Blue indicates a high probability density.

Fig. 1: Cluster visualisation and heatmap matrix.

3 Interface design

Our interface consists of: (1) A *hierarchically-clustered* diagram for clutter management, (2) a *heatmap matrix* for navigating between dimensions, (3) a *graphical history* for navigation context, and (4) a *word cloud* for inspecting clusters.

3.1 Hierarchically-clustered scatterplot

We construct a cluster hierarchy, allowing fluid navigation between multiple levels. This supports *exploratory* analysis, where it is not known in advance which aspects of the data are most important. We use agglomerative hierarchical clustering using Euclidean distance and centroid linkage. Every datapoint is initialised as a separate 'cluster'. In each iteration, the pair of clusters with the lowest inter-cluster distance is merged. This is repeated until all points have been merged into a single cluster. This process creates a tree (represented as a dendrogram (Fig. 3, left)): the root node is a cluster containing the whole dataset, nodes have exactly two children, and leaves are individual datapoints.

Visualising a cluster Clustering trades detail about individual data points for aggregate information. A good cluster representation would convey information about its contents (*scenting*) for effective exploration (*foraging*) [17]. In our representation of each cluster (Fig. 1a), the *shape* of the cluster is preserved by rendering the convex hull of its constituent points; the colour of a cluster is mapped to its *cardinality* – darker clusters contain more points; and, the centroid is plotted in red. Data points (words) are shown explicitly, and can be inspected individually, if a cluster contains very few of them.

Cluster expansion Double-clicking a cluster expands it, ideally resulting in a display that efficiently uses the available screen space while minimising overlap of the newly displayed clusters. Each expansion may correspond to a descent of multiple levels in the hierarchy tree, based on a criterion that supports the fastest descent of the hierarchy while avoiding clutter.

We developed the heuristic of a 'minimal displayable centroid distance'. The idea is that the centroids of clusters onscreen should never be closer than this amount. We set this to 30 pixels, corresponding roughly to 1cm on our displays. Clusters

Fig. 2: Word clouds corresponding to four clusters. Font size and colour encode the words' distances from the centroid. Can you assign a meaning to each cloud?

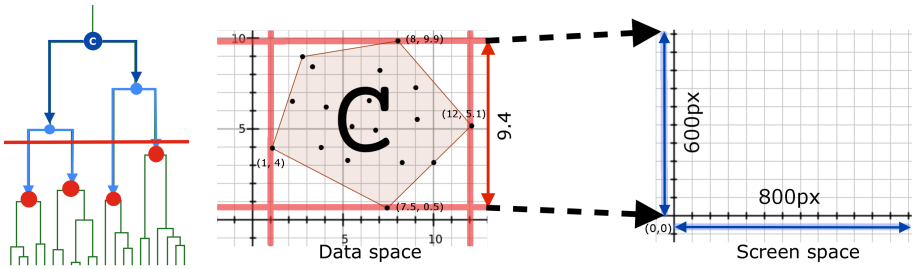


Fig. 3: Expanding a cluster. When C is expanded, a cut (red line) is made at the height corresponding to the minimum displayable distance between clusters.

higher in the hierarchy tree correspond to a greater centroid distance between that cluster's children. On expanding a cluster, the scatterplot is rescaled to tightly fit the expanded cluster, such that the minimum and maximum value on the x and y axes corresponds to the extents of the cluster on the dimensions being plotted on x and y, respectively. The pixel size of the overall scatterplot is known, giving a mapping between data and screen space. From this, we map a distance of 30 pixels back to data space and get the optimal tree cut height (the lowest height where clusters are sufficiently distant) (Fig. 3).

3.2 Heatmap matrix: helping users navigate between dimensions

The user must select which two dimensions of the n-dimensional dataset will be plotted. Without guidance, this task can degenerate into tedious enumeration of dimension pairs, or ineffective random switching. A scatterplot matrix displays all dimension pairs, letting users quickly identify plots of interest, but is costly to render: for 30,000 words it requires plotting 30,000 points per dimension pair. One strategy to reduce the rendering cost is to display a naïve random sample, but this only works on uniformly distributed data; with outliers and areas of varying density, it produces distorted or misleading plots.

Our solution is to plot the *sampled probability density* of the data as a heatmap, with colour mapped to density, as seen in Fig. 1b. We used bivariate kernel density estimation (KDE) [18]. This significantly reduces the complexity of rendering while still capturing the overall shape of the data. The resultant heatmap matrix is a navigational aid: users click on cells in this matrix to select which two dimensions are displayed in the cluster diagram.

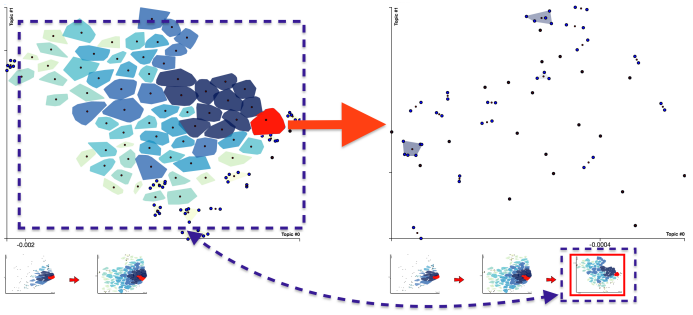


Fig. 4: Cluster expansion step. Expanded clusters are marked in red.

3.3 Word cloud

There are two ways to inspect words. Hovering on single points displays a tooltip that remains open if the point is clicked. When a cluster is clicked, a subset of the words contained in the cluster is visualised in a word cloud to the right. To manage the cloud's visual complexity, only the 30 words closest to the cluster centroid are displayed, as these are most representative of the cluster. The size and colour of words are mapped to the distance of the words from the centroid.

3.4 Graphical expansion history

When a cluster is expanded, its place within the larger cluster hierarchy is no longer visible. A graphical history [19] preserves this context. Upon cluster expansion, a snapshot of the current plot, highlighting the expanded cluster, is added to the history. A sequence of expansions provides context for each expanded cluster, showing how the expanded cluster relates both to its immediate context as well as the entire data space (Fig. 4). Any snapshot in the history can be clicked to revert to that level, creating a multi-level overview+detail [20].

Taken together, the four components: clustered scatterplot, heatmap matrix, graphical history, and word cloud constitute our interface (Fig. 5). The heatmap matrix is accessed with the 'change dimensions' button, which displays the matrix to the right of the cluster diagram in place of the word cloud.

4 User study

We define two goals of latent semantic space exploration: (1) finding groups of related words and assigning a meaning to the common underlying theme, and (2) interpreting the meaning of each dimension. We were interested in evaluating:

- **Effectiveness**: were the two goals of exploration achieved?
- **Style**: was exploration *broad*, exploring many combinations of dimensions, or *deep*, emphasising word inspection and navigation within dimension pairs?
- **Usability**: do users find the system usable?

We conducted an experiment to assess these questions, comparing our interface with the following two alternatives. Firstly, a *plain scatterplot* system replaces the cluster visualisation with a scatterplot that users can pan and zoom – a

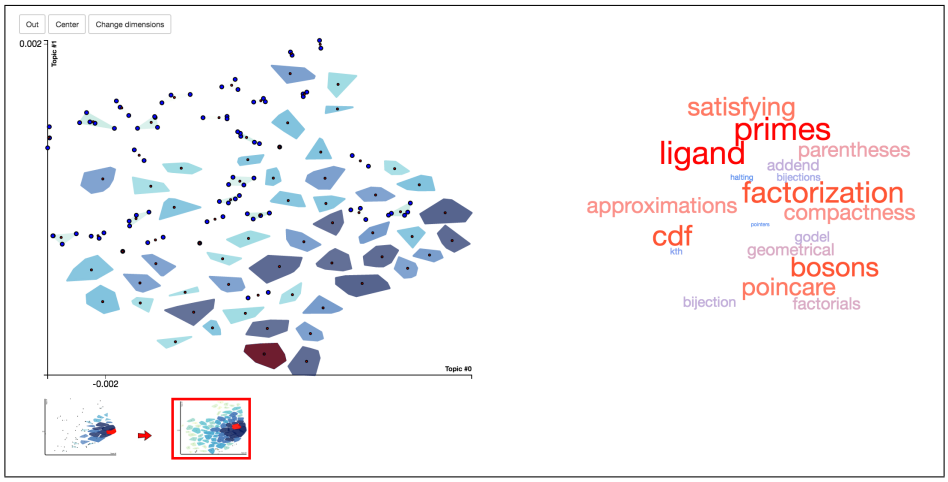
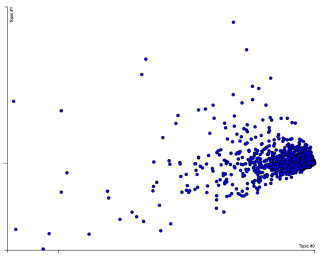
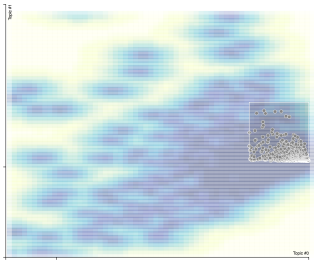


Fig. 5: Our interface in use.



(a) Plain scatterplot with pan/zoom.



(b) Heatmap showing selected points.

Fig. 6: Experimental alternative interfaces

conventional clutter management strategy (Fig. 6a). Secondly, a *heatmap* system uses the KDE heatmap as the primary display (Fig. 6b). To view individual points, the user selects an area of the plot, within which individual points are rendered. When the selection is made, the plot zooms into the selection, and individual points are rendered which can be inspected using tooltips. Both systems retain the navigation matrix, but lose the graphical history and word cloud, leaving tooltips as the method for word inspection.

4.1 Experiment design and procedure

Twelve Cambridge University undergraduates with no prior exposure to LSA were recruited via convenience sampling. The experiment was a within-subjects comparison of the three systems. Participants were briefed on LSA and on the operation of each interface. For each of the three interfaces, participants carried out an experimental task, then completed two usability questionnaires.

In each task, the participant was instructed to (i) assign a meaning to groups of words which they found to be related, and (ii) assign a meaning to each of

the dimensions. Three disjoint datasets of 30,000 words were sampled from the corpus. The design was counterbalanced: each dataset was assigned to each interface in equal representation across participants. The 3 systems were presented to participants in different orders, each of the 6 possible orders being assigned exactly twice. These measures mitigated learning effects and order effects.

We recorded the number of meanings offered by the user, counting at most one assigned meaning per word group / dimension, even if the participant offered multiple interpretations. Participants were free to continue exploration as long as they desired. General usability ratings were obtained using IBM's Post-Study System Usability Questionnaire (PSSUQ), while IBM's After-Scenario Questionnaire (ASQ) was used to measure task-specific usability [21]. Both use a 7-point scale with lower values reflecting superior usability.

4.2 Results

We refer to our interactive Cluster diagram as C, the Heatmap alternative as H, and the plain Scatterplot alternative as S. All post-hoc tests were subjected to Bonferroni correction.

Assignment of meaning: Participants assigned meaning to significantly more word groups using C (average of 7.92 groups) versus H (5.33 word groups, $p = 0.037$) and S (4.25, $p = 0.038$). (Planned contrasts after one-way repeated measures ANOVA yielded $F(2, 22) = 5.162$, $p = 0.019$). A significant difference was not found in the number of meaning assignments for dimensions.

Style of exploration: We studied how often users switched the dimensions displayed using the heatmap matrix. Participants switched dimensions several times in S, but less frequently when using C. In contrast, participants inspected a far greater number of words with C than with either alternative. C therefore promoted a more depth-first style of exploration due to the ease of navigating the hierarchy, facilitating model interpretation grounded in specific words. Concretely: a significant Friedman's test was followed with Wilcoxon signed rank tests. Users switched dimensions more often with S ($p = 0.037$). With a similar analysis, the number of words inspected was significantly different ($p = 0.028$). The average number of words explored when using C was 1517 ($p = 0.010$), as compared to 479 with H and 772 with S ($p = 0.050$).

Usability: Users found C more usable than either alternative. C significantly improved the users' exploration effectiveness in terms of the number of groups of related words found. This was expected, as the word cloud allows more words to be inspected simultaneously, and clusters encapsulate the semantics of a given word group. Concretely: The PSSUQ score for C (average 1.86) was significantly better than H (average 4.08, $p = 0.002$) or S (average 3.01, $p = 0.015$) (Wilcoxon signed-rank tests following significant Friedman's test ($p = 0.001$)). The differences in task-specific ASQ ratings for the *dimension interpretation* task were significant ($p = 0.002$). C was rated better than both S (mean difference -1.389, $p = 0.02$), and H (mean difference -1.528, $p = 0.001$). For the *word group* task we observed similar, but non-significant mean differences.

5 Conclusion

Latent semantic spaces are a valuable tool for the analysis of large text corpora. However, interpreting latent semantic spaces is difficult, and visual scalability is a major design challenge, as is accessibility for non-experts.

Our novel interface uses a hierarchical clustering approach to clutter reduction, allowing users to gain an overview of semantic structure in the corpus. The cluster diagram can be combined with summary distributions arranged in a heatmap matrix. A user study showed that the usability of our interactive diagram was significantly superior to alternatives based on either plain scatterplots or heatmaps alone. Moreover, the hierarchical cluster diagram facilitated the identification and assignment of meaning to more word groups.

References

1. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse Processes **25**(October) (1998) 259–284
2. Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research **37** (2010) 141–188
3. Landauer, T.K., Laham, D., Rehder, B., Schreiner, M.E.: How Well Can Passage Meaning be Derived without Using Word Order? A Comparison of Latent Semantic Analysis and Humans. Proc. Cognitive Science Society (1997) 412–417
4. Landauer, T.K., Laham, D., Derr, M.: From paragraph to graph: latent semantic analysis for information visualization. Proc. NAS USA **101** (2004) 5214–5219
5. Habernal, I., Brychcín, T.: Semantic spaces for sentiment analysis. LNCS **8082 LNAI** (2013) 484–491
6. Elmqvist, N., Dragicevic, P., Fekete, J.D.: Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation. IEEE TVCG **14**(6) (2008) 1141–1148
7. Fua, Y.H., Ward, M.O., Rundensteiner, E.A.: Hierarchical parallel coordinates for exploration of large datasets. Proc. Vis. (1999) 43–508
8. Carr, D.B., Littlefield, R.J., Nicholson, W.L., Littlefield, J.S.: Scattterplot Matrix Techniques for Large N. J. American Stat. Assoc. **82**(398) (1987) 424–436
9. Zhu, W., Chen, C.: Storylines: Visual exploration and analysis in latent semantic spaces. Computers & Graphics **31** (2007) 338–349
10. Ellis, G., Dix, A.: A Taxonomy of Clutter Reduction for Information Visualisation. **13**(6) (2007) 1216–1223
11. Elmqvist, N., Fekete, J.D.: Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. IEEE Transactions on Visualization and Computer Graphics **16**(3) (2010) 439–454
12. Shaoul, C. and Westbury C.: The Westbury Lab Wikipedia Corpus, Edmonton, AB: University of Alberta (2010)
13. Radim ehek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, ELRA (may 2010) 45–50
14. Baker, K.: Singular value decomposition tutorial. Ohio State University **24** (2005)
15. Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W.: Handbook of Latent Semantic Analysis. (2007)
16. Salvador, S., Chan, P.: Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. 2012 IEEE 24th International Conference on Tools with Artificial Intelligence **0** (2004) 576–584
17. Pirolli, P., Card, S.: Information foraging. Psychological review **106**(4) (1999) 643
18. Wang, M.P., Jones, M.C.: Comparison of smoothing parameterization in bivariate kernel density estimation. J. American Statistical Assoc. **88**(422) (1993) 520–528
19. Kurlander, D., Feiner, S.: Editable graphical histories. In: IEEE Visual Languages. (1988) 127–134
20. Cockburn, A., Karlson, A., Bederson, B.B.: A review of overview+detail, zooming, and focus+context interfaces. ACM Comput. Surv. **41** (2008) 1–31
21. Lewis, J.: IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. IJHCI **7**(1) (1995) 57–78