Demonstrating Conceptual Dynamics in an Evolving Text Collection

Sándor Darányi and Peter Wittek
University of Borås

Abstract

Based on real world user demands, we demonstrate how animated visualisation of evolving text corpora displays the underlying dynamics of semantic content. To interpret the results, one needs a dynamic theory of word meaning. We suggest that conceptual dynamics as the interaction between kinds of intellectual, emotional etc. content, and language, is key for such a theory. We demonstrate our methodology by two-way seriation which is a popular technique to analyse groups of similar instances and their features, as well as the connections between the groups themselves. The two-way seriated data may be visualised as a two-dimensional heat map or as a three-dimensional landscape where colour codes or height correspond to the values in the matrix. In this paper we focus on two-way seriation of sparse data in the Reuters-21568 test collection. To achieve a meaningful visualisation thereof we introduce a compactly supported convolution kernel similar to filter kernels used in image reconstruction and geostatistics. This filter populates the high-dimensional sparse space with values that interpolate nearby elements, and provides insight into the clustering structure. We also extend two-way seriation to deal with online updates of both the row and column spaces, and, combined with the convolution kernel, demonstrate a three-dimensional visualisation of dynamics.

1 Introduction

Apart from handling the temporality of symbolic content as a more and more pronounced challenge (Bosi and Ragot, 2012; Hashemi et al., 2012), our motivation for the research introduced below goes back to a certain increasing demand for time-dependent semantic content visualisation in digital document collections (Kumar et al., 1998; Krishnan and Jones, 2005; Van de Sompel et al., 2009; Sanderson and Van de Sompel, 2012).To exemplify real world user needs, an archive may be interested in an institutional celebrity index to measure the spread and acceptance of certain ideologies over time, or a historical map of protests in a region, or cross-reporting between allied countries on political dissent. Apart from targeted tool development for e.g. mapping literary scandals about plagiarism as the only visible form of contesting cultural monopolies in a regime, or visualising social toxicities by e.g. pinpointing changes in "criminality", "moral decay", and "alcoholism", or the shift of power within a bloc by following such index terms as "gaps", "polemics", "controversy", "difference", "conflict", "reactions" in order to establish the changing poles of ideological focus and influence, the application areas for such temporal drilling into semantic content are seemingly endless. The question is how to construct semantic spaces which enable one to observe and exploit the tectonics of ideas.

The fact that geometric models are suitable for information retrieval (IR) and text categorization (TC) suggests that geometry, prominently vector spaces, can carry meaning (Widdows, 2004). Because there exist geometries commuting with probability (Birkhoff and von Neumann, 1936; van Rijsbergen, 2004; Khrennikov, 2010), our argumentation below holds for a very general case. All this notwithstanding the fact that vector space in its most basic form is not semantic, although the ghost in the machine, i.e. its ability to yield results which nevertheless make sense, goes back to the fact that the context of sentence content is partially preserved even after having eliminated stopwords which are of limited or no use for indexing. This means that Harris' distributional hypothesis holds (Harris, 1970), indirectly supported by Wittgenstein's contextual theory of meaning ("meaning is use"). As a result, vector space contains semantics even if it is not evident to the naked eye. This is exploited by more advanced vector based indexing and retrieval models such as Latent Semantic Analysis (Deerwester et al., 1990) and Random Indexing (Kanerva et al., 2000)

From the perspective of language models (as the concept is being used in IR research where it is associated with a document in a collection and related to a query as its input so that retrieved documents are ranked based on the probability that the document's language model would generate the terms of the query), the workings of the mind are hierarchical because the mapping between conceptual and semantic content is constrained by syntactic relations, i.e.

the mapping is many-to-few (Yuret, 1998). Therefore we regard meaning as a layered product of the mind due to whose capacity for abstraction, conceptual structures have a strong hierarchical component. Layered meaning needs a layered vehicle: such conceptual structures map onto words, sentences, passages, with motifs, topics etc. hinting at intermediate strata of unspecified positions. Likewise, if it is not uttered phonemic sequences or written symbol strings that carry concepts, geometry has to do the same favour instead.

But in which sense is geometry stratified? On a basic level, similarity as a measure is continuous: there are no two locations in a geometry whose similarity of content cannot be measured by the similarity of those vectors pointing at them. Secondly, mapping as an operation is continuous as well: there are no two locations in a geometry which could not be mapped onto each other. These two layers of continuity enable one to map semantic content on some of the locations while others remain empty, resulting in vector space representations of such content (including probabilistic models which depart from binary vectors). The outcome is a vehicle whose two levels, similarity and mappings, can carry units of semantic content layered in its own right, although more often than not passage or document level representation is indicated by the presence and absence of text words only, devoid of grammar (the so-called bag-of-words [BOW] model).

We explicitly distinguish between fields used as mathematical support for reasoning by modal logic (Resconi et al., 2000) and semantic fields (Trier, 1934; Lehrer, 1975) or lexical fields (Cruse, 1986) of individual words known from linguistics and their reconstructions from word-word matrices in semantic space models such as HAL (Lund and Burgess, 1996), versus topic models of documents (Blei et al., 2003). Although both of the latter are applied to describe regions of reasonably uniform semantic content, they also clearly do so in two consecutive magnitudes. We also note in passing that one cannot disprove the continuity of document topics, or prove that there exist locations or regions in vector space where by some default content cannot be mapped.

Semantic content mapped onto vector space coordinates is modelled either as exactly located, which is the majority approach, or as inexactly (regionally) located, an observation from cognitive psychology (Erk, 2009). We must consider that both approaches describe complementary aspects of truth (Darányi and Wittek, 2012a) and thereby face an observable whose two possible and mutually exclusive outcomes cannot be measured with arbitrary precision, i.e. are subject to Heisenberg's uncertainty relation. Another such pair, precision vs. recall in IR and TC evaluation was suggested by (Dominich, 2001, p.144). Certainly, imprecise but complete versus precise but incomplete retrieved sets hint at the workings of the same constraint, plus both observables yield results according to the design of the observation apparatus. All these indications point at the need to come up with more comprehensive models of semantic content whose nature may be closer to representations in complex vector space (Zuccon et al., 2011; De Vine and Bruza, 2010). Such Hilbert spaces, as the reference to Heisenberg had already implied, happen to be home ground for quantum mechanics and quantum-like models of, among others, language (Aerts and Czachor, 2004; Khrennikov, 2010).

From this emerging, more complete description, in this paper we want to focus on the dynamics of semantic content as inherent in language change (Baker, 2008). On the one hand, linguistics has been developing new theories to this end (Veltman, 1996; van Eijck and Visser, 2010). On the other hand, as practical information science knows only all too well, due to language change, indexing terminology of documents keeps on undergoing modifications. Consider the following examples for terminology change over time: gramophone - record player; computer screen - monitor; duplicating machine - photocopier; European Coal and Steel Community - Common Market - European Economic Community or EEC - European Community or EC - European Union or EU; bicycle - bike - BMX. These are just a few of many more which make it an issue that in order to index and retrieve documents from certain periods, the right language must be used.

Such changes in terminology are handled by vocabulary control and ontology building, but - and this is our key argument here - unless the dynamic nature of evolving semantic content is addressed, understood and modelled, its adequate management will always be out of reach. Therefore we suggest to look at this very dynamics and want to visualise it as the tectonics of continental plates in geology. The tool for this will be evolving heat maps and evolving landscapes.

One way to model a continuum of ideas underlying terminology change is to consider it sampled by thinking and the result mapped to a discretised representation, i.e. language. In this metaphor, by the time it appears in language, conceptual content is carried by space-delimited uttered and/or written vehicles such as words and expression forms above them, a view ultimately going back to (Hägerstrand, 1953) as (Persson and Ellegård, 2012) point out, the use of such trajectories illustrated e.g. by (Wilson, 2008). This will be our working assumption, consequently we apply an annealing methodology to "reconstruct" the original, hypothetical form and its dynamics. Our method will be to replace biclustering or block clustering (Hartigan, 1975) as the simplest way to create three-dimensional landscapes from semantic content by biseriation, which may be commonplace for archaeology (Peeples and Schachner, 2012) and

bioinformatics (Caraux and Pinloche, 2005), but less so for digital libraries.

This paper is organized as follows. In Section 2, we briefly discuss previous research from a subject area perspective and offer a definition of conceptual dynamics. We discuss related semantic regionality in Section 3. In Section 4, we depart from existing algorithmic alternatives to present seriation, two-way seriation and dynamic seriation as consecutive modifications of the underlying solution. Section 5 lists our results including the material and methods used, with Section 6 summing up our conclusions and pointing at future research.

2 Previous Research

To measure structural changes over time in science (White and McCain, 1998; Chen and Carr, 1999; Shiffrin and Börner, 2004; Börner et al., 2006), or in bibliographic citation patterns (Chen, 2004, p.316) (Nelson, 2006) has been a matter of interest to several research communities in information science since the eighties. Visualising such processes by taking snapshots of their phases has lately become a pronounced field of activity in information visualisation research (Mane and Börner, 2004). However, this focus of interest has not been explicitly connected with language change (Baker, 2008) in general, or conceptual changes in vocabulary control in particular. The latter is important for two reasons: the possibility of changes in vocabulary control as a means of indexing in databases raises the question, how to find the right term at the right time at the right place, i.e. the spatiotemporal stratification of indexing and search terminology. Secondly, the underlying mathematical framework is far from being finalized or fully understood. We argue that connecting semantics and dynamics by modelling the former on the latter yields new insight in the tectonics of concept evolution, a process that can be visualised e.g. by the cartographic metaphor (Marcus, 1994; Eppler and Burkhard, 2004). We note in passing that while similar research into the representation of semantic content as a landscape predates ours (Wise et al., 1995; Wise, 1999), the present approach goes back to the metaphor of modelling language change on classical mechanics (Darányi and Wittek, 2012a,b).

Recognizing news in texts by topic detection and tracking (Allan et al., 1998) and new event or burst detection (Papka, 1999) is another relevant research direction, in essence similar to time series analysis. Significant solutions range from extracting time-varying features from texts (Swan and Allan, 1999) to constructing timelines for event classification based on word usage statistics (Swan and Jensen, 2000) and personalized newsfeeds based on information novelty (Gabrilovich et al., 2004). In the latter, the inter- and intra-document dynamics of documents is considered to model how information evolves over time from article to article, as well as within individual articles. Such methods can be applied to the analysis of temporal dynamics in online text streams such as newsfeed or e-mail (Kleinberg, 2003, 2006), or chronologically ordered documents (Fung et al., 2005).

Ongoing IR research fruitfully applies the above to the identification and tracking of changing user interests for personalized information systems (Crabtree and Soltysiak, 1998), going back to measuring the number of elementary topic shifts within XML documents based on topic segmentation (Ashoori et al., 2007). Starting with identifying new events in event-based IR (Papka, 1999), accomplishments include a time-dependent language model to address the temporality of relevance (Li and Croft, 2003), and using timestamps on queries to assess average precision on the basis of time-dependent probability distributions (Diaz and Jones, 2004). In a similar vein, Krause et al. (2006) have come up with a unified model of topic intensity tracking combined with document classification.

A very close parallel to conceptual dynamics discussed the visualisation of thematic changes in documents over time based on the river metaphor (Havre et al., 2000). Another example for static distributions of semantic content is (Minghim et al., 2005) for fast content-based visual mapping of document collections. As such static content maps are produced by a great variety of methods for information visualisation, we suggest that as a next step, one needs to convert them into animated sequences to study their inherent dynamics.

2.1 Definition

Socioeconomic, technological, political etc. changes are in reciprocal relationship with conceptual development which, once reflected in language, results in indexing terminology change, in turn affecting the quality of IR and TC in particular, and access to digital objects in general. We refer to this continuum of value fluctuations as conceptual dynamics (CD) and define it as follows: CD as a phenomenon is relevant to the semantic and ontological continuity and comparability of collection content, including the selection, preservation, maintenance, collection and archival of digital assets. In this paper it will stand for the exploration of topical continuities and/or discontinuities over spatiotemporal regions.

In a technical sense, CD is a visual approach to demonstrate the changes in the semantic landscape in a growing collection of documents. As new documents are added, the existing distributional pattern of words and expressions changes and their relations and relative importance shift. This leads to a dynamic pattern that we wish to visualise to gain insight into evolving concepts.

3 The Regionality of Semantic Content

Whereas in vector models of TC and IR it is a tacit assumption that semantic content is exactly located by position vectors pointing at term, document and query meanings, referring to cognitive theories, recently (Hoenkamp, 2011) has suggested that concepts also constitute a vector space of their own. This begs one to consider linguistic parallels of the same theories, in contrast arguing for a regional nature of semantic content which seems to be our common psychological heritage (Shepard, 1987) and can become part of a vector space model as well.

In linguistics, the regional nature of word semantics can be best observed on the overlap between word senses displayed as semantic fields (Lehrer, 1975; Dyvik, 2005). Priss and Old (2005) model the underlying, language-independent conceptual regions by neighbourhood lattices. Further the very concern itself is not new, IR and TC having had assumed for a long time that the immediate neighbourhood of relevant terms and documents contains related, and therefore important, information, which can be used for e.g. relevance feedback (Rocchio, 1971). Moreover there is an argument in (Widdows, 2008) about support vector machines (SVM) linked to quantum disjunctions, the link being regions, meant to solve the problem to be able to say that apple is a kind of fruit (apple is part of the fruit region, i.e. its hyponym), as opposed to modelling that apple and fruit have something to do with each other. SVMs do this by finding the separating hyperplane, but more research is needed to understand whether the separating hyperplane indeed defines a region. Finally, instead of regions, Zuccon et al. (2009) measure the distance between subspaces spanned by documents by projecting them into one another.

Reinforcing Dyvik's and Priss and Old's argument, Erk also argues for the regionality of word meaning, i.e. its inexact location (Erk, 2009). She departs from the fact that many models of categorization in psychology represent a concept as a region, characterized by feature vectors with dimension weights, and offers two computational models for monosemous and polysemous words, both of which can host soft region boundaries. Using so-called type vectors as central vectors, each type vector comes with a vector which defines the importance of each dimension, thus the type vector and its weight vector define a region. Here, regionality implies gradually decreasing similarity between document, query and term vectors.

Regionality also manifests itself if term vectors are embedded into an L2 space, assigning sums of sinusoids or wavelets to each term in the function space (Wittek and Darányi, 2007; Wittek and Tan, 2011). In these models the length of the period or the length of the support controls the inexactness of semantic content, and given that terms are arranged according to a semantic order, this representation may lead to improvement in classification performance. Interference (pattern) models using functions to represent semantic content implement this regionality expectation (Azzopardi, 2000; Dorrer et al., 2001)

4 Mapping Evolving Semantic Structures

Whereas the above section described relevant previous research, to introduce our selected solution, we also mention that several algorithms exist for the mapping of evolving semantic content. For instance, self-organizing maps (SOMs) have been used to provide a visual overview of the clusters in expanding document collections (Kohonen et al., 2000). A SOM is a two-dimensional grid of a neural network in which the nodes are adjusted as subsequent document vectors stimulate the network. Eventually clusters of documents will be assigned to nodes and the overall structure can be visualised. If instead of document vectors, term vectors are used to train the network, we may view the resulting map as a form of CD.

In the neural network of a SOM, the two-dimensional layout is somewhat arbitrary and only relative positions matter. The x and y axes are not meaningful. The axes in a matrix layout, however, are typically easily interpreted. For instance, the row space may refer to features and the column space to instances that the features describe. In a term-document matrix the features are index terms (terms that are semantically charged), whereas instances are documents (a text, a web page, a blog entry, etc.). If we rearrange the rows and columns in the matrix such that related terms will be near each other, and clusters of similar documents will also be nearby column vectors, we arrive at a visual reflection of a statistical model, a heat map of the matrix (Wilkinson and Friendly, 2009). The underlying idea of a heat map is a

seriation of both the row and the column space. We regard seriation as "sequencing objects along a continuum that rely upon a symmetric proximity measure defined between the objects to be seriated" (Hubert, 1974). Seriation is widely used in the visualisation of binary matrices (Chen, 2002; Liiv et al., 2011).

Seriation is a combinatorial data analysis method that reorders instances into a sequence along a one-dimensional continuum. The basis of the reordering is a pairwise distance between the instances. Seriation algorithms place instances next to each other if the distance between them is small, and the eventual order reveals regularity and patterns in the whole series (Liiv, 2010). Seriation is different from clustering: clustering groups nearby objects together, but it does not necessarily reveal relations among the groups. Yet both approaches are NP hard problems (Wilkinson and Wills, 2005, p.525).

Seriation as described above is uncommon. Two-way seriation is the typical application. Two-way seriation assumes a matrix representation, or a two-dimensional layout of the data, where rows correspond to instances and columns correspond to features that describe the individual instances (this assignment of rows and columns is arbitrary and can happen the other way). This approach has a history of over a hundred years (see (Liiv, 2010) for an overview), and it is commonly used in a wide variety of information visualisation methods, including microarray data (Eisen et al., 1998; Caraux and Pinloche, 2005), binary matrices (Liiv et al., 2011), and others (Chen, 2002). Thus two-way seriation introduces a kind of regionality: a local neighbourhood is guaranteed to have related elements. Seriation is similar to clustering, and two-way seriation is similar to biclustering: instances with similar feature subsets are grouped together in regular patterns. The difference is that the overall structure of the seriated two-dimensional array is meaningful, the groups follow each other in an optimized order. In other words the overall structure is not an arbitrary ordering of row and column cluster trees. Certain types of two-way seriation are also called heat maps (Wilkinson and Friendly, 2009).

In the context of text mining, Wittek et al. (2009) introduced a seriation that used a distance function that did not only rely on the distributional patterns of words, but also included an external lexical database that encoded word relations. This method translated seriation to a graph problem in which an approximate solution to a minimum-weight Hamiltonian path was sought (Rosenkrantz et al., 1977). The weights in the path corresponded to semantic distances, and the nodes to words. We rely on this algorithm, with seriation extended to both the column and row space. We also added the option for dynamic updates, inserting new elements into the existing order with a greedy approach.

Two-way semantic seriation of a dense matrix will result in a heat map, that is, a two-dimensional grid of coloured squares where the hot spots will correspond to closely related terms and documents. While sparse data ask for similar seriation methods, visualising the result is harder. (Berry et al., 1996) tested different methods to improve the browsing experience of hypertext documents with promising results, although the plots of the sparse matrices resulted from document indexing were not immediately useful for visual analysis. Since the vast majority of the values are zero, a heat map or a three-dimensional view of it would not be meaningful. A similar problem is present in information retrieval and text classification: similar documents or documents corresponding to queries are not necessarily found if they do not share the exact same index terms. The trick to deal with the problem is to introduce a kind of 'smoothing' that fills the vector space with non-zeroes if the location is relevant in the corresponding document or query. Common methods include dimensionality reduction techniques such as singular value decomposition (Deerwester et al., 1990) and random projection (Kanerva et al., 2000).

A different technique took the one-way seriation of feature space as its starting point and introduced compactly supported wavelets over the feature space (Wittek and Tan, 2011). The distance function between the documents (or between the documents and the queries) measured the overlap between the wavelets. This is similar to image filtering in two dimensions, where convolution is performed on the pixels with a kernel. It is a commonly used operation in blurring, sharpening, edge detection, and other image processing steps. The Gaussian filter (also known as Gaussian blur or the two-dimensional Weierstrass transform) convolves the data with the two-dimensional Gaussian function $G(x, y) = \frac{1}{2\pi(\sigma_x^2+\sigma_y^2)} \exp(-\frac{x^2+y^2}{2(\sigma_x^2+\sigma_y^2}))$. This transform can be viewed as a 'smoothed' version of the data: a value of the transform at the point (x, y) is obtained by averaging the values of the data, weighted with a Gaussian centred at (x, y). To perform the convolution on discrete data, a convolution matrix is generated, typically with dimensions $6\sigma_x \times 6\sigma_y$, eventually leading to a compactly supported kernel. Gaussian filters have been used in language modelling to address data sparseness (Chen and Rosenfeld, 1999). We introduce this compactly supported two-dimensional Gaussian filter over the two-way seriated sparse data. This approach technically embeds the data into the $L_2(\mathbb{R}^2)$ space while 'blurring' nearby items together. This process is also similar to normalized convolution used in sparse image reconstruction when the kernel is a Gaussian (Pham and van Vliet, 2003; Foster and Evans, 2008). If the data is assumed to be Gaussian, ordinary kriging also yields similar results (Boucher et al., 2006).

The colour values of a heat map can be interpreted as a height coordinate, and, using two-dimensional wavelets as Parzen window estimators over the points, we may derive a three-dimensional surface visualisation of the matrix.

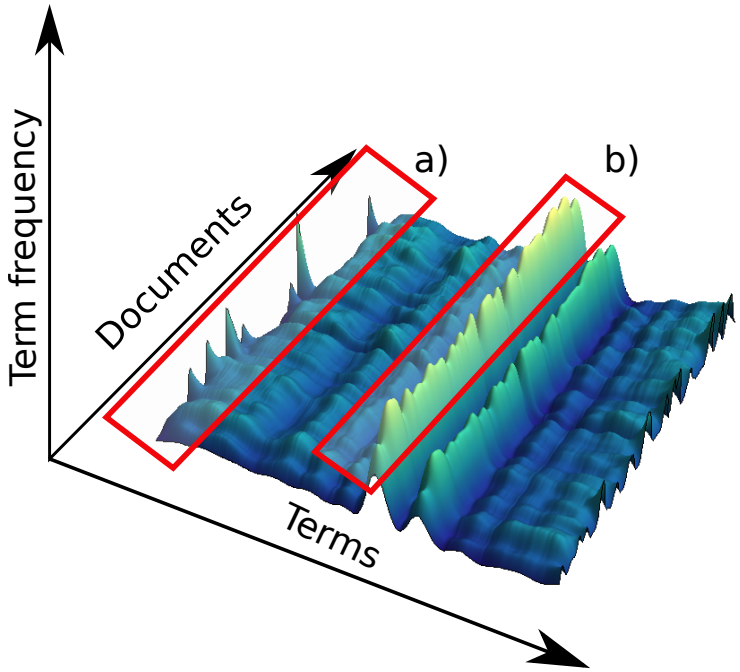


Figure 1: Interpreting a three-dimensional landscape. Rectangle a) highlights an index term that is distant from the other terms; it is very common in some documents, but absent in most. Rectangle b) highlights another index term which is commonly found in every document.

Such methods have been used in the past to visualise the topic structure of document collections, also applying seriation (Wise et al., 1995; Davidson et al., 1998). Extensions of these methods have also been developed to analyse gene expressions (Stuart et al., 2003). The ability to capture the temporal changes of a collection is somewhat limited in these methods, whereas a different, two-dimensional visualisation method focuses exclusively on dynamics (Havre et al., 2000). Figure 1 provides some insights to aid the interpretation of the three-dimensional landscape of the document collection described in the next section.

Note that in our visualisation, points on the x-y axes can have labels: in the case of textual data, they are the index terms and the documents. The landscape that emerges in three dimensions does not represent explicit information of the concepts that underlie the nearby index terms or documents. We will not be able to pinpoint a location and declare that this position is anything beyond a specific word form referring to an underlying concept. The labelling problem persists just like in automated topic modelling.

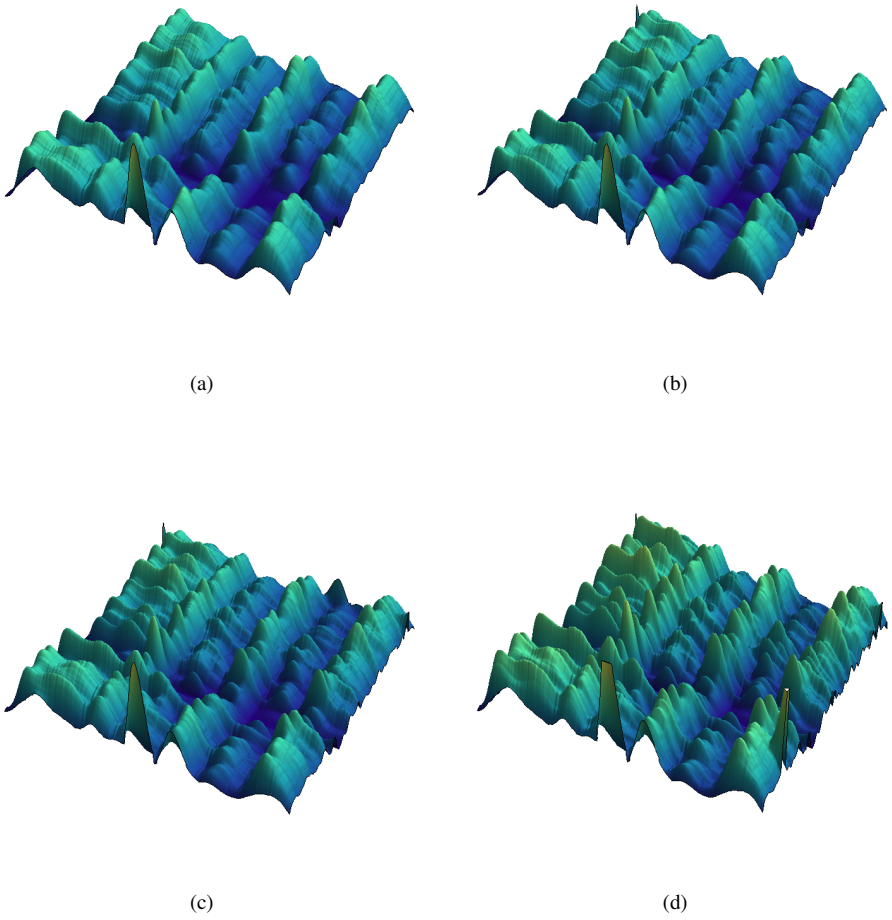
5 Material, Methods and Results

5.1 Collection and Preprocessing

The collection consists of the subtitles of 21,578 news reports that appeared on the Reuters wire between February and October 1987. We filtered words that appear less than three times. Stop words were removed and stemming was performed by the SnowBall English stemmer. All index terms were converted to lower case. Numbers were discarded. Indexing was performed with Lucene 3.6.0. The resulting index had 10,074 index terms. We used tf-idf weighting to smooth the space. We trained an initial two-way seriation with 9,000 documents, and then inserted new documents one by one to derive the dynamic maps.

5.2 Results and Discussion

Figure 2 demonstrates the dynamic updates of two-way seriation. Starting with one hundred documents as the basis of the left-right seriation (Subfigure 2(a)), eventually the insert heuristic expands the seriation to the full collection



(a) (b)

(c) (d)

Figure 2: Four snapshots of the dynamic data with approximately three thousand new instances in each step

(Subfigure 2(d)).

The above prove that animated visualisation of evolving text corpora displays the underlying dynamics of semantic content. However, given that text collections more often than not are embodied in sparse matrices whereas their visual output, according to our methodology, is continuous, this discrepancy needs an explanation.

On the one hand, our argumentation takes into consideration that the representation of document content is the result of multiple samplings: first, semantic content in the mind, i.e. concepts need to be mapped to language, then, language to mathematical objects. Whereas it is beyond our competence to tell if concepts exist in some kind of continuity in the mind or are discrete, the first mapping is definitely a discretisation process, therefore we decided to give priority to a continuous-to-discrete mapping of semantics from mind to language in our current working hypothesis.

On the other hand, contrasting the view that word meaning, and thereby also document and query content, are exactly localized in vector space, we cited counter-arguments for the regional nature of semantic content. Recently Darányi and Wittek (2012a) suggested that this duality may be essential, with neither options being exclusive. In this sense the regional aspect of word meaning could be envisaged as a discretised variant of continuous concepts in one's thinking, a reflection by the necessity of sampling to turn concepts into language.

To interpret results from our working hypothesis, one needs a dynamic theory of word meaning. Apart from linguistics where recently such proposals have started to appear, we suggest that conceptual dynamics as the interaction between kinds of semantic, affective, functional, aesthetic etc. content, and language, is an indicator and in support of such a theory.

6 Conclusions and Future Research

Changes in the semantic content of data may affect a collection time and again and result in discontinuities of terminology, leading to access problems. Whereas worldwide, the handling of this temporal aspect of collections is becoming a research topic in its own right (Van de Sompel et al., 2009; Bosi and Ragot, 2012; Hashemi et al., 2012), a comprehensive frame of thought to understand the interlinked linguistic and documentation-related issues is a necessary step toward problem solving. To this end, we have introduced the notion of conceptual dynamics and its technological implementation to show how, by large-scale two-way seriation of sparse data, temporal updates of both column and row vectors can be seamlessly incorporated. By adding kernel-based filtering for better visualisation of sparse data in 3-D, and applying plate tectonics as a visualisation metaphor, we demonstrated considerable semantic modifications over time in the composition of the Reuters-21578 test collection. This indicates that (1) conceptual dynamics in one's mental space or mind can find a discretised expression in texts, (2) such processes affect the composition of text collections, and finally (3) by exploiting the regional nature of word meaning, one can approximate the original conceptual continuity. Secondly, we consider the more explicit modelling of sense relations between terms an important direction for future research. However, in order to add more shades of meaning and a pragmatic aspect to index terms, important for the advanced communication of content, the currently used mathematical objects used for representing such content must be swapped for types with a higher representation capacity, e.g. vectors for functions. Thirdly, we plan to test and evaluate the scalability and interpretability aspects of our results.

We plan to address some parallelization issues of biseriation for conceptual dynamics before scalability experiments can start. In this respect, an expansion of our efforts is considered by calling in (Benoit, 2002) to add data discretization for the identification of clusters of related semantic content. We anticipate that approaches measuring and utilizing conceptual dynamics can anchor the concept of meaning in IR and thereby pave the way for a more cognitive understanding of indexing, storage and retrieval.

7 Acknowledgements

We thank two unknown reviewers for their detailed and thoughtful suggestions which resulted in an improved version of this manuscript. We are also are pleased to acknowledge an exemplification of archive-related user needs by Gabriella Ivacs and Ioana Macrea (Central European University, Hungary). Further the first author is grateful to Johan Eklund (University of Borås, Sweden) for discussions on the subject.

References

Aerts, D. and Czachor, M. (2004). Quantum aspects of semantic analysis and symbolic artificial intelligence. *Journal of Physics A: Mathematical and General*, 37:L123–L132.

Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. (1998). Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*.

Ashoori, E., Lalmas, M., and Tsikrika, T. (2007). Examining topic shifts in content-oriented XML retrieval. *International Journal on Digital Libraries*, 8(1):39–60.

Azzopardi, L. (2000). Wave motion: A new metaphor for 2d information visualization the exploration of a metaphor. Technical report, University of Newcastle.

Baker, A. (2008). Computational approaches to the study of language change. *Language and Linguistics Compass*, 2(3):289–307.

Benoit, G. (2002). Data discretization for novel relationship discovery in information retrieval. *Journal of the American Society for Information Science and Technology*, 53(9):736–746.

Berry, M., Hendrickson, B., and Raghavan, P. (1996). Sparse matrix reordering schemes for browsing hypertext. *Lectures in Applied Mathematics*, 32:99–124.

Birkhoff, G. and von Neumann, J. (1936). The logic of quantum mechanics. *The Annals of Mathematics*, 37(4):823–843.

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Börner, K., Penumarthy, S., Meiss, M., and Ke, W. (2006). Mapping the diffusion of scholarly knowledge among major US research institutions. *Scientometrics*, 68(3):415–426.

Bosi, S. and Ragot, L. (2012). Time representation in economics. *Theoretical Economics Letters*, 2(1):10–15.

Boucher, A., Seto, K., and Journel, A. (2006). A novel method for mapping land cover changes: Incorporating time and space with geostatistics. *IEEE Transactions on Geoscience and Remote Sensing*, 44(11):3427–3435.

Caraux, G. and Pinloche, S. (2005). PermutMatrix: a graphical environment to arrange gene expression profiles in optimal linear order. *Bioinformatics*, 21(7):1280–1281.

Chen, C. (2002). Generalized association plots: Information visualization via iteratively generated correlation matrices. *Statistica Sinica*, 12(1):7–30.

Chen, C. (2004). *Information visualization: Beyond the horizon*. Springer-Verlag.

Chen, C. and Carr, L. (1999). Trailblazing the literature of hypertext: Author co-citation analysis. In *Proceedings of HT-99, 10th Conference on Hypertext and Hypermedia*, pages 51–60, Darmstadt, Germany.

Chen, S. and Rosenfeld, R. (1999). A Gaussian prior for smoothing maximum entropy models. Technical report, Carnegie Mellon University.

Crabtree, B. I. and Soltysiak, S. J. (1998). Identifying and tracking changing interests. *International Journal on Digital Libraries*, 2(1):38–53.

Cruse, D. (1986). *Lexical semantics*. Cambridge University Press, New York, NY, USA.

Daránvi, S. and Wittek, P. (2012a). Connecting the dots: Mass, energy, word meaning, and particle-wave duality. In *Proceedings of QI-12, 6th International Quantum Interaction Symposium*, Paris, France.

Daránvi, S. and Wittek, P. (2012b). The gravity of meaning: Physics as a metaphor to model semantic changes. In *Proceedings of SLTC-12, 4th Swedish Language Technology Conference*, Lund, Sweden.

Davidson, G., Hendrickson, B., Johnson, D., Meyers, C., and Wylie, B. (1998). Knowledge mining with VxInsight: Discovery through interaction. *Journal of Intelligent Information Systems*, 11(3):259–285.

De Vine, L. and Bruza, P. (2010). Semantic oscillations: Encoding context and structure in complex valued holographic vectors. In *Proceedings of QI-10, 4th Symposium on Quantum Informatics for Cognitive, Social, and Semantic Processes*, pages 11–13, Arlington, VA, USA.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Diaz, F. and Jones, R. (2004). Using temporal profiles of queries for precision prediction. In *Proceedings of SIGIR-04, 27th Annual International Conference on Research and Development in Information Retrieval*, pages 18–24, Sheffield, United Kingdom. ACM.

Dominich, S. (2001). *Mathematical Foundations of Information Retrieval*. Kluwer Academic Publishers Norwell, MA, USA.

Dorrer, C., Londero, P., Anderson, M., Wallentowitz, S., and Walmsley, I. (2001). Computing with interference: all-optical single-query 50-element database search. In *Proceedings of QELS-01, Quantum Electronics and Laser Science Conference*, pages 149–150.

Dyvik, H. (2005). Translations as a semantic knowledge source. In *Proceedings of HLT-05, 2nd Baltic Conference on Human Language Technologies*, pages 27–38, Tallinn, Estonia.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868.

Eppler, M. and Burkhard, R. (2004). Knowledge visualization: towards a new discipline and its fields of application. In *Encyclopedia of Knowledge Management*.

Erk, K. (2009). Representing words as regions in vector space. In *Proceedings of CoNLL-09, 13th Conference on Computational Natural Language Learning*, pages 57–65, Boulder, CO, USA.

Foster, M. and Evans, A. (2008). Performance evaluation of multivariate interpolation methods for scattered data in geoscience applications. In *Proceedings of IGARSS-08, International Geoscience and Remote Sensing Symposium*, volume 4, Boston, MA, USA.

Fung, G., Yu, J., Yu, P., and Lu, H. (2005). Parameter free bursty events detection in text streams. In *Proceedings of VLDB-05, 31st International Conference on Very Large Data Bases*, pages 181–192, Trondheim, Norway.

Gabrilovich, E., Dumais, S., and Horvitz, E. (2004). Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In *Proceedings of WWW-04, 13th International Conference on the World Wide Web*, pages 482–490, New York City, NY, USA.

Hägerstrand, T. (1953). *Innovationsförloppet ur korologisk synpunkt [Innovation diffusion as a spatial process]*. Gleerup, Lund, Sweden.

Harris, Z. (1970). Distributional structure. In Harris, Z., editor, *Papers in structural and transformational Linguistics*, Formal Linguistics, pages 775–794. Humanities Press, New York, NY, USA.

Hartigan, J. (1975). *Clustering algorithms*. Wiley New York.

Hashemi, F., Hongler, M., and Gallay, O. (2012). Spatio-temporal patterns for a generalized innovation diffusion model. *Theoretical Economics Letters*, 2(1):1–9.

Havre, S., Hetzler, B., and Nowell, L. (2000). ThemeRiver: Visualizing theme changes over time. In *Proceedings of Infovis-00, IEEE Symposium on Information Visualization*, pages 115–123, Salt Lake City, UT, USA.

Hoenkamp, E. (2011). Trading spaces: on the lore and limitations of latent semantic analysis. In *Proceedings of ICTIR-11, 3rd International Conference on the Theory of Information Retrieval*, pages 40–51, Bertinoro, Italy.

Hubert, L. (1974). Some applications of graph theory and related non-metric techniques to problems of approximate seriation: The case of symmetric proximity measures. *British Journal of Mathematical and Statistical Psychology*, 27(2):133–153.

Kanerva, P., Kristofersson, J., and Holst, A. (2000). Random indexing of text samples for latent semantic analysis. In *Proceedings of CogSci-00, 22nd Annual Conference of the Cognitive Science Society*, volume 1036, Philadelphia, PA, USA.

Khrennikov, A. (2010). *Ubiquitous quantum structure: from psychology to finance*. Springer Verlag.

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397.

Kleinberg, J. (2006). Temporal dynamics of on-line information streams. *Data Stream Management: Processing High-Speed Data Streams*.

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., and Saarela, A. (2000). Self organization of a massive text document collection. *IEEE Transactions on Neural Networks*, 11(3):574–585.

Krause, A., Leskovec, J., and Guestrin, C. (2006). Data association for topic intensity tracking. In *Proceedings of ICML-06, 23rd International Conference on Machine Learning*, pages 497–504, Pittsburgh, PA, USA. ACM.

Krishnan, A. and Jones, S. (2005). Timespace: activity-based temporal visualisation of personal information spaces. *Personal and Ubiquitous Computing*, 9(1):46–65.

Kumar, V., Furuta, R., and Allen, R. (1998). Metadata visualization for digital libraries: interactive timeline editing and review. In *Proceedings of DL-98, 3rd ACM Conference on Digital libraries*, pages 126–133, Pittsburgh, PA, USA.

Lehrer, A. (1975). *Semantic fields and lexical structure*. American Elsevier, New York, NY, US.

Li, X. and Croft, W. (2003). Time-based language models. In *Proceedings of CIKM-03, 12th International Conference on Information and Knowledge Management*, pages 469–475, New Orleans, LA, USA.

Liiv, I. (2010). Seriation and matrix reordering methods: An historical overview. *Statistical Analysis and Data Mining*, 3(2):70–91.

Liiv, I., Opik, R., Ubi, J., and Stasko, J. (2011). Visual matrix explorer for collaborative seriation. *Wiley Interdisciplinary Reviews: Computational Statistics*.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods Instruments and Computers*, 28:203–208.

Mane, K. and Börner, K. (2004). Mapping topics and topic bursts in PNAS. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5287–5290.

Marcus, A. (1994). Managing metaphors for advanced user interfaces. In *Proceedings of AVI-94, 2nd Workshop on Advanced Visual Interfaces*, pages 12–18, Bari, Italy.

Minghim, R., Paulovich, F., and de Andrade Lopes, A. (2005). Fast content-based visual mapping for interactive exploration of document collections. Technical report, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Nelson, M. (2006). Visualization of citation patterns of some Canadian journals. *Scientometrics*, 67(2):279–289.

Papka, R. (1999). *On-line new event detection, clustering, and tracking*. PhD thesis, University of Massachusetts Amherst.

Peeples, M. and Schachner, G. (2012). Refining correspondence analysis-based ceramic seriation of regional data sets. *Journal of Archaeological Science*, 39(8):2818–2827.

Persson, O. and Ellegård, K. (2012). Torsten hägerstrand in the citation time web. *The Professional Geographer*, 64(2):250–261.

Pham, T. and van Vliet, L. (2003). Normalized averaging using adaptive applicability functions with applications in image reconstruction from sparsely and randomly sampled data. In *Proceedings of SCIA-03, 13th Scandinavian Conference on Image Analysis*, pages 485–492, Halmstad, Sweden.

Priss, U. and Old, L. (2005). Conceptual exploration of semantic mirrors. In *Proceedings of ICFCA-05, 3rd International Conference on Formal Concept Analysis*, Lens, France.

Resconi, G., Mura, T., and Shimbo, M. (2000). Semantic field and accessibility relations. In *Proceedings of KES-00, 4th International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies*, volume 2, pages 491–494, Wellington, New Zealand.

Rocchio, J. (1971). Relevance feedback in information retrieval. In Salton, G., editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Rosenkrantz, D., Stearns, R., Lewis, P., et al. (1977). An analysis of several heuristics for the traveling salesman problem. *SIAM Journal on Computing*, 6(3):563–581.

Sanderson, R. and Van de Sompel, H. (2012). Cool URIs and dynamic data. *Internet Computing*, 16(4):76–79.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317.

Shiffrin, R. and Börner, K. (2004). Mapping knowledge domains. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5183–5185.

Stuart, J., Segal, E., Koller, D., and Kim, S. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255.

Swan, R. and Allan, J. (1999). Extracting significant time varying features from text. In *Proceedings of CIKM-99, 8th International Conference on Information and Knowledge Management*, pages 38–45, Kansas City, MO, USA.

Swan, R. and Jensen, D. (2000). Timemines: Constructing timelines with statistical models of word usage. In *Proceedings of KDD-2000 Workshop on Text Mining*, pages 73–80, Boston, MA, USA.

Trier, J. (1934). Das sprachliche feld. *Neue Jahrbucher fur Wissenschaft und Jugendbildung*, 10:428–449.

Van de Sompel, H., Nelson, M., Sanderson, R., Balakireva, L., Ainsworth, S., and Shankar, H. (2009). Memento: Time travel for the web. *Arxiv preprint arxiv:0911.1112*.

van Eijck, J. and Visser, A. (2010). Dynamic semantics. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

van Rijsbergen, C. J. (2004). *The Geometry of Information Retrieval*. Cambridge University Press, New York, NY, USA.

Veltman, F. (1996). Defaults in update semantics. *Journal of Philosophical Logic*, 25(3):221–261.

White, H. and McCain, K. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science*, 49(4):327–355.

Widdows, D. (2004). *Geometry and meaning*. CLSI Publications.

Widdows, D. (2008). Semantic vector products: Some initial investigations. In *Proceedings of QI-08, 2nd International Symposium on Quantum Interaction*, Oxford, UK.

Wilkinson, L. and Friendly, M. (2009). The history of the cluster heat map. *The American Statistician*, 63(2):179–184.

Wilkinson, L. and Wills, G. (2005). *The grammar of graphics*. Springer Verlag.

Wilson, C. (2008). Activity patterns in space and time: calculating representative hagerstrand trajectories. *Transportation*, 35(4):485–499.

Wise, J. (1999). The ecological approach to text visualization. *Journal of the American Society for Information Science*, 50(13):1224–1233.

Wise, J., Thomas, J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., and Crow, V. (1995). Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *Proceedings of Infovis-95, IEEE Symposium on Information Visualization*, volume 51, Atlanta, GA, USA.

Wittek, P. and Darányi, S. (2007). Representing word semantics for IR by continuous functions. In Dominich, S. and Kiss, F., editors, *Studies in Theory of Information Retrieval. Proceedings of ICTIR-07, 1st International Conference of the Theory of Information Retrieval*, pages 149–155, Budapest, Hungary.

Wittek, P., Darányi, S., and Tan, C. L. (2009). An ordering of terms based on semantic relatedness. In Bunt, H., editor, *Proceedings of IWCS-09, 8th International Conference on Computational Semantics*, Tilburg, The Netherlands.

Wittek, P. and Tan, C. L. (2011). Compactly supported basis functions as support vector kernels for classification. *Transactions on Pattern Analysis and Machine Intelligence*, 33(10):2039 –2050.

Yuret, D. (1998). Discovery of linguistic relations using lexical attraction. *Arxiv preprint cmp-lg/9805009*.

Zuccon, G., Azzopardi, L., and Rijsbergen, C. (2009). Semantic spaces: Measuring the distance between different subspaces. In *Proceedings of QI-09, 3rd International Symposium on Quantum Interaction*, pages 225–236, Saarbruecken, Germany.

Zuccon, G., Piwowarski, B., and Azzopardi, L. (2011). On the use of complex numbers in quantum models for information retrieval. In *Proceedings of ICTIR-11, 3rd International Conference on the Theory of Information Retrieval*, Bertinoro, Italy.