A Physical Metaphor to Study Semantic Drift

Sándor Daránya
Swedish School of Library and Information Science
University of Borås
Borås 50190, Sweden
sandor.daranyi@hb.se

Peter Wittek
Swedish School of Library and Information Science
University of Borås
Borås 50190, Sweden
ICFO-The Institute of Photonic Sciences
Castelldefels 08860, Spain

Konstantinos Konstantinidis
Information Technologies Institute
The Centre for Research & Technology, Hellas
Thessaloniki 57001, Greece
konkonst@iti.gr

Symeon Papadopoulos
Information Technologies Institute
The Centre for Research & Technology, Hellas
Thessaloniki 57001, Greece
papadop@iti.gr

Efstratios Kontopoulos
Information Technologies Institute
The Centre for Research & Technology, Hellas
Thessaloniki 57001, Greece
skontopo@iti.gr

ABSTRACT

In accessibility tests for digital preservation, over time we experience drifts of localized and labelled content in statistical models of evolving semantics represented as a vector field. This articulates the need to detect, measure, interpret and model outcomes of knowledge dynamics. To this end we employ a high-performance machine learning algorithm for the training of extremely large emergent self-organizing maps for exploratory data analysis. The working hypothesis we present here is that the dynamics of semantic drifts can be modeled on a relaxed version of Newtonian mechanics called social mechanics. By using term distances as a measure of semantic relatedness vs. their PageRank values indicating social importance and applied as variable 'term mass', gravitation as a metaphor to express changes in the semantic content of a vector field lends a new perspective for experimentation. From 'term gravitation' over time, one can compute its generating potential whose fluctuations manifest modifications in pairwise term similarity vs. social importance, thereby updating Osgood's semantic differential. The dataset examined is the public catalog metadata of Tate Galleries, London.

CCS Concepts

•**Computing methodologies** → **Lexical semantics;** *Neural networks;* •**Information systems** → Similarity measures;

Keywords

Semantic drift; vector field semantics; emergent self-organizing maps; content dynamics; gravitational model.

1. INTRODUCTION

The evolving nature of digital collections comes with an extra difficulty: due to various but constant influences inherent in updates, the interpretability of the data keeps on changing. This manifests itself as concept drift [47] or semantic drift [49, 16], the gradual change of a concept's semantic value as it is perceived by a community. Despite terminology differences, the problem is real and with the increasing scale of digital collections, its importance is expected to grow [37]. If we add drifts in cultural values as well, the fallout from their combination brings memory institutions in a vulnerable position as regards long term digital preservation. We illustrate this on a museum example, the subject index of the Tate Galleries, London. In our example, semantic drifts lead to limited access by Information Retrieval (IR). The methodology we apply to demonstrate our point is vector field semantics by emergent self-organizing maps (ESOM) [44], because the interpretation of semantic drift needs a theory of update semantics [46], integrated with a vector field rather than a vector space representation of content [50, 49]. Further, given such content dynamics, we argue that for its modeling, one can fall back on tested concepts from classical (Newtonian) mechanics and differential geometry. For such a framework, e.g. similarity between objects or features can be considered an attractive force, and changes over time manifest in content drifts have a quasi-physical explanation. The main contributions of this paper are the following:

1. A methodology for the detection, measurement and interpretation of semantic drift;

2. On drift examples, an improved understanding of how semantic content as a vector field 'behaves' over time by falling back on physics as a metaphor;

3. As a consequence of the above, the concept of semantic potential as a combined measure of semantic relatedness and semantic importance.

2. BACKGROUND

2.1 Terminology

Evolving semantics (also often referred to as 'semantic change' [42]) is an active and growing area of research into language change [5] that observes and measures the phenomenon of changes in the meaning of concepts within knowledge representation models, along with their potential replacement by other meanings over time. Therefore it can have drastic consequences on the use of knowledge representation models in applications. Semantic change relates to various lines of research such as ontology change, evolution, management and versioning [24], but it also entails ambiguous terms of slightly different meanings, interchanging shifts with drifts and versioning, and applied to concepts, semantics and topics, always related to the thematic composition of collections [54, 45, 20]. A related term is semantic decay as a metric: it has been empirically shown that the more a concept is reused, the less semantically rich it becomes [27]. Though largely counter-intuitive, this derivation is based on the fact that frequent usage of terms in diverse domains leads to relaxing the initially strict semantics related to them. The opposite would hold if a term was persistently used within a single domain (or in to a great extent similar domains), which would lead to its gradual specialization and enrichment of its semantics.

2.2 Related Research

Here we mention four relevant directions, all of them contributors to our understanding of a complex issue in their overlap.

2.2.1 Temporality and Advanced Access

By advanced access to digital collections we mean the spectrum of automatic indexing, automatic classification, IR, and information visualization. All of the aforementioned can have a temporal aspect: trend analysis, emergence of concepts or ideas, representation of the past and the future, network dynamic, shaping and decay of communities, and in general, any Web research topic where a dynamic understanding is superior to a static view, requires integration of the time dimension. Examples comprise e.g the presentation, organization and exploration of search results [3] in the context of web dynamics and analytics including the dynamics of user behaviour [32]; interacting with ephemeral content of the historical web [1], visualizing the evolution of image content tags [13], or temporal topic detection without citation analysis [38]. A related but separate research area for the above is in the overlap of cultural heritage and IR [22, 11].

2.2.2 Vector Space vs. Vector Field Semantics

For an IR model to be successful, its relationship with at least one major theory of word meaning has to be demonstrated. With no such connection, meaning in numbers becomes the puzzle of the ghost in the machine. For the vector space IR model (VSM) - underlying many of today's competitive IR products and services - such a connection can be demonstrated; for others like PageRank [7], the link between graph theory and linear algebra leads to the same interpretation. Namely, in both cases, the theory of word semantics cross-pollinating numbers with meaning is of a contextual kind, formalized by the distributional hypothesis [17] which posits that words occurring in similar contexts tend to have similar meanings. As a result, the respective models can imitate the field-like continuity of conceptual content. However, unless we consider the VSM roots of both the probabilistic relevance model[1] and its spinoffs including BM25,[2] such a link is still waiting to be shown between probability and semantics [15].

Although several attempts exist to this end [41, 31], a brief overview should be helpful. Looking for a good fit with some reasonably formalized theory of semantics, two immediate questions emerge. First, can the observed features be regarded as entries in a vocabulary? If so, distributional semantics applies and, given more complex representations, other types may do so as well [52]. The second question is, do they form sentences? For example, one could regard a workflow (process) a sentence, in which case compositional semantics applies [8, 35]. If not, theories of word semantics should be considered only. Below we shall depart from this assumption.

Notwithstanding the fact that vector space in its most basic form is not semantic, its ability to yield results which make sense goes back to the fact that the context of sentence content is partially preserved even after having eliminated stop-words which are useless for document indexing. This means that Wittgenstein's contextual theory of meaning ('Meaning is use') holds [53], also pronounced by the distributional hypothesis. This is exploited by more advanced vector based indexing and retrieval models such as Latent Semantic Analysis (LSA) [12] or random indexing [19], as well as by neural language models, ranging from the Simple Recurrent Networks, and their very popular flavour, Long Short-Term Memory [18], or the recently proposed Global Vector for Word Representation [29], which are currently considered to be the state-of-the-art approach for text representation. However, we should also remember another approach paraphrased as 'Meaning is change', namely the stimulus-response theory of meaning proposed e.g. by Bloomfield[3] in anthropological linguistics and Morris [4] in behavioral semiotics, plus the biological theory of meaning [43]. These authors stress that the meaning of an action is in its consequences. Consequently word semantics should be represented not as a vector space with position vectors only, but as a dynamic vector field with both position and direction vectors [50].

2.2.3 Linguistic 'Forces'

As White suggests, linguistics, like physics, has four binding forces [48]:

1. The strong nuclear force, which is the strongest 'glue' in physics, corresponds to word uninterruptability (binding morphemes into words);

2. Electromagnetism, which is less strong, corresponds to grammar and binds words into sentences;

[1] Because it departs from a 'binary index descriptions of documents', see [34].
[2] See p. 339 in [33].
[3] en.wikipedia.org/wiki/Leonard_Bloomfield
[4] en.wikipedia.org/wiki/Charles_W._Morris

3. The weak nuclear force, being even less strong, compares to texture or cohesion (also called coherence), binding sentences into texts;

4. Finally gravity as the weakest force acts like intercohesion or intercoherence which binds texts into literatures (i.e. documents into collections or databases).

Mainstream linguistics traditionally deals with Forces 1 and 2, while discourse analysis and text linguistics are particularly concerned with Force 3. The field most identified with the study of Force 4 is information science. As the concept of force implies, referring here to attraction, it takes energy to keep things together, therefore the energy doing so is stored in agglomerations of observables of different kinds in different magnitudes, and can be released from such structures. A notable difference between physical and linguistic systems is that extracting work content, i.e. 'energy' from symbols by reading or copying them does not annihilate symbolic content. Looking now at the same problem from another angle, in the above and related efforts, 'energy' inherent in all four types can be the model of e.g. a Type 2, i.e. electromagnetism-like attractive-repulsive binding force such as lexical attraction, also known as syntactic word affinity [6] or sentence cohesion, such as by modeling dependency grammar by mutual information [55]. In a text categorization and/or IR setting, a similar phenomenon is term dependence based on their co-occurrence.

2.2.4 Semantic Kernels and 'Gravity'

A radial basis function (RBF) kernel, being an exponentially decaying feature transformation, has the capacity to generate a potential surface and hence create the impression of gravity, providing one with distance-based decay of interaction strength, plus a scalar scaling factor for the interaction, i.e. $K(x, x') = exp(-||x - x'||^2)$ [25]. We know that semantic kernels and the metric tensor are related, hence some kind of a functional equivalent of gravitation shapes the curvature of classification space [4, 14]. At the same time, gravitation as a classification paradigm [28] or a clustering principle [2] is considered as a model for certain symptoms of content behavior.

3. WORKING HYPOTHESIS & METHODOLOGY

In order to combine semantics from computational linguistics with evolution, we select the theory of semantic fields [40] and blend it with multivariate statistics plus the concept of fields in classical mechanics to bring it closer to Veltman's update semantics [46], and to enable machine learning. Our working hypothesis for experiment design is as follows:

- Semantic drifts can be modeled on an evolving vector field as suggested by [49, 50];

- To follow up on the analogy from semantic kernels defining the curvature of classification space and let this curvature evolve, Newton's universal law of gravitation can be adapted to the idea of the dynamic library [36]. To this end, we model similarity by $F = Gm_1m_2/r^2$, with term dislocations over epochs stored in distance matrices. Ignoring G, we shall use the PageRank value of index terms on their respective hierarchical levels for mass values. Since force is the negative gradient of potential, i.e. $F(x) = -dU/dx$, we can compute this potential surface over the respective term sets to conceptualize the driving mechanism of semantic drifts;

- The potential following from the gravity model manifests two kinds of interaction between entries in the indexing vocabulary of a collection. Over time, changes in collection composition lead to different proportions of semantic similarity vs. authenticity between term pairs, expressed as a cohesive force between features and/or objects.

3.1 ESOMs and Somoclu

3.1.1 Vector Field Creation by ESOMs

In the various flavours of the VSM, we work with an $m \times n$ matrix in which columns are indexed by documents and rows by terms. We shall focus here on the m term vectors only, which identify specific locations in the n-dimensional space spanned by the documents.

A scalar or vector field is defined at all points in space, so it is insufficient to have a value at the discrete locations identified by the term vectors. To assign a vector value to each point in space, we work on a two-dimensional surface. All term vectors have a location on this surface. All the other points on the surface which do not have a vector assigned to them are interpolated.

The assignment of points on the surface and the term vectors is done by training a self-organizing map, that is, a grid of artificial neurons. Each node in the grid is associated with a weight vector of n dimensions, matching the term vectors. Taking a term vector, we search for the closest weight vector, and pull it slightly closer to the term vector, repeating the procedure with the weight vectors of the neighboring neurons, with decreasing weight as we get further away from the best matching unit. Then we take the next term vector and repeat this from finding the best matching unit until every term vector is processed. We call a training round that uses all term vectors an epoch. We can have subsequent training epochs with a smaller neighborhood radius and a lower learning rate. While there is no criterion for a convergence, we can continue training epochs until the topology of the network no longer shows major changes. The resulting map reflects the local topology of the original high-dimensional space [21].

Since we would like to train large maps to get a meaningful approximation in the space between term vectors, we turn to a high-performance implementation called Somoclu[5] [51].

3.1.2 Drift Detection

The task of drift detection, measurement and interpretation is carried out in three basic steps as follows:

- Step 1: Somoclu maps the high-dimensional topology of multivariate data to a low-dimensional (2-d) embedding by ESOM. The algorithm is initialized by LSA, Principal Component Analysis (PCA), or random indexing, and creates a vector field over a rectangular grid of nodes of an artificial neural network, adding

[5] https://peterwittek.github.io/somoclu/

continuity by interpolation among grid nodes. Due to this interpolation, content is mapped onto those nodes of the neural network that represent best matching units (BMUs).

- Step 2: Clustering over this low-dimensional topology marks up the cluster boundaries to which BMUs belong. Their clusters are located within ridges or watersheds [44, 39, 23]. Content splitting tendencies are indicated by the ridge wall width and height around such basins so that the method yields an overlay of two aligned contour maps in change, i.e. content structure vs. tension structure. In Somoclu, nine clustering methods are available. Because self-organizing maps, including ESOM, reproduce the local but not the global topology of data, the clusters should be locally meaningful and consistent on a neighborhood level only.

- Step 3: Evolving cluster interpretation by semantic consistency check can be measured relative to an anchor (non-shifting) term used as the origin of the 2-d coordinate system, or by distance changes from a cluster centroid, etc. In parallel, to support semiautomatic evaluation, variable cluster content can be expressed for comparison by histograms, pie diagrams, or other visualization methods.

4. DATASET AND EXPERIMENT DESIGN

4.1 Tate Subject Index

Tate holds the national collection of British art from 1500 to the present day and international modern and contemporary art. The collection embraces all media, from painting, drawing, sculpture and prints to photography, video and film, installation and performance. The 19th century holdings are dominated by the Turner Bequest with cca 30,000 works of art on paper, including watercolors, drawings and 300 oil paintings. The catalog metadata for the 69,202 artworks that Tate owns or jointly owns with the National Galleries of Scotland are available in JSON format as open data.[6] Out of the above, 53,698 records are timestamped. The artefacts are indexed by Tate's own hierarchical subject index which has three levels, from general to specific index terms.[7]

4.2 Analysis Framework Description

To study the robust core of a dynamically changing indexing vocabulary, we filtered the dataset for a start. As statistics for the Tate holdings show two acquisition peaks in 1796-1844 (33,625 artworks) and 1960-2009 (12,756 artworks), we focused on these two periods broken down into 10 five-years epochs each, with altogether 46,381 artworks. In the 19th century period, subject index level 1 had 22 unique general index terms (21 of them persistent over ten epochs), level 2 had 203 unique intermediate index terms (142 of them persistent), and level 3 had 6624 unique specific index terms (225 of them persistent). In the 20th century period, level 1 had 24 unique terms (22 of them persistent), level 2 used 211 unique terms (177 of them persistent), and level 3 had 7536 unique terms (288 of them persistent over ten epochs).

[6]github.com/tategallery/collection

[7]tate.org.uk/art/artworks/turner-self-portrait-n00458

Table 1: Sample index terms describing a Turner self-portrait

level 1 (general)	level 2 (intermediate)	level 3 (specific)
Objects	Clothing and personal effects	Cravat
People	Adults	Man
Named individuals	Turner, Joseph Mallord William	-
Portraits	Self-portraits	-
Work and occupations	Arts and entertainment	Artist, painter

Table 1 displays a sample entry from the subject index. Following text pre-processing, which included the application of tokenization and stop-word removal on all three levels of concepts in the subject index, adjacency matrices and subsequently graphs were created using the co-occurrence of the terms in the artworks as undirected, weighted edges. These matrices were then used to extract an importance measure for each term by employing the PageRank algorithm, and to create ESOM maps using the Somoclu implementation.

For each of the 80 epochs (2 periods x 4 levels x 10 epochs), the ESOM's codebook was first initialized by employing PCA with randomized SVD, which was then used for mapping the high-dimensional co-occurrence data to an ESOM with a toroid topology. The results were represented on the two-dimensional projection of the toroid using different granularities according to the indexing level (20x12 = level 1, 40x24 = level 2, 50x30 = level 3, 60x40 = all levels together). Introducing the least displaced term per indexing level over a period as an anchor against which all term drifts on that level could be measured, we tracked the tension vs. content structure of evolving term semantics and evaluated the resulting term clusters for their semantic consistency.

The input matrices were processed by Somoclu as described above and the codebook of each ESOM was clustered using the affinity propagation algorithm. The results were tested for robustness by hierarchical cluster analysis (HCA), using Euclidean distance as similarity measure and farthest neighbor (complete) linkage to maximize distance between clusters, keeping them thereby both distinct and coherent. The ESOM-based cluster maps expressed the evolving semantics of the collection as a series of 2-dimensional landscapes over 10 epochs times two periods.

Term drift detection, measurement and interpretation were based on these maps. To enable drift measurement, we generated a parallel set of maps with the term of greatest importance over all periods as its anchor point. Importance was defined by the Reciprocal Rank Fusion coefficient [10] which combined the PageRank values of each term over all periods. This relative location was used for the computation of respective term-term distance matrices over every epoch of each period. Term dislocations over epochs were logged, recording both the splits of term clusters mapped onto a single grid node in a previous epoch, or the merger of two formally independent nodes labelled with different terms into a single one. These splits and merges were used to define the drift rate and subsequently the stability of the lexical field.

Finally, as per the second point of the working hypothesis,

the gravity and potential surfaces for every epoch were computed. When computing gravity and potential, the property of mass was expressed via each term's PageRank score and the distance by measuring the normalized (sum to 1) Euclidean distance between the corresponding BMU vectors.

5. RESULTS

Index term drift detection, measurement and evaluation were based on the analysis of ESOM maps, leading to drift logs on all indexing levels. Parallel to that, covering every time step of collection development, we also extracted normalized histograms to describe the evolving topical composition of the collection, and respective pie charts to describe the thematic composition of the clusters. Further, to check cluster robustness, HCA dendrograms were computed for term-term matrices, also compared with those from term-document matrices. On one hand, these gave us a detailed overview of semantic drift in the analyzed periods. On the other hand, the observed dynamics could be modeled on the gravitational force and its generating potential.

A more detailed report would go beyond the opportunities of this paper. However, some key indications were the following.

5.1 Semantic Drifts

Content mapping means that term membership for every cluster in every time step is recorded and term positions and dislocations over time with regard to an anchor position are computed, thereby recording the evolving distance structure of indexing terminology. This amounts to drift detection and its exact measurement. Adding a drift log results in extracted lists of index terms on all indexing hierarchy levels plus their percentage contrasted with the totals. Drifts can be partitioned into splits and merges. In case of a split, two concept labels that used to be mapped on the same grid node in one epoch become separated and tag two nodes in the next phase, while for a merge, the opposite holds. From an IR perspective splits decrease recall and merges decrease precision, limiting the quality of access; from the perspective of long term digital preservation, they indicate at-risk indexing terminology. Splits and merges were listed by Somoclu for every epoch over both periods. For instance a sample semantic drift log file recorded that due to new entries in the catalog in 1796-1800, by 1800 on subject index level 2, for drifting words i.e. 'art', 'works', 'scientific', 'measuring', 'monuments', 'places', 'workspaces'. Therefore, based on the same subject index terms, anyone using this tool in 1800 would have been unable to retrieve the same objects as in 1796. In a vector field, all the terms and their respective semantic tags are in constant flux due to external social pressures, such as e.g. new topics over items in the collection due to the composition of donations or fashion. Without data about these pressures quasi embedding and shaping the Tate collection, the correlations between social factors and semantic composition of the collection could not be explicitly computed and named. Still, some trends could be visually recognized over both series of maps, going back to their relatively constant semantic structure where temporary content dislocations did not seriously disturb the relationships between terms, i.e. neighboring labels tended to stick with one another, such as 'towns, cities, villages' vs. 'inland' and 'natural'. In other words, the lexical fields as locally represented by Somoclu remained relatively stable.

The stability of these fields was measured in terms of drift rates which were computed by detecting the splits and merges that happened to the BMUs (e.g. 1). Specifically, we were not looking at the distance they travelled, rather at the fact that they formed or joined or moved away from a cluster (i.e. a BMU) in between epochs.

Overall, in this particular collection, splits between level 1 concepts took place occasionally, whereas both splits and merges occurred on indexing levels 2-3 on a regular basis. The drift rate was increasingly high: for level 2 index terms, it was 19-22 % in the 1796-1845 period vs. 15-27.5 % in 1960-2009, whereas for level 3 terms it was 29-57 % (1796-1845) vs. 54-61 % (1960-2009). These percentages suggest that the more specific the subject index becomes, the more volatile its terminology, especially with regard to modern art.

5.2 Content vs. Tension Structure

To describe the composition of the social tensions shaping this collection, one can compare e.g. the level 2 indexing vocabularies for both periods. In general, this is where one witnesses the workings of language change, part producing new concepts, part letting certain index terms decay. E.g. focus is shifting from a concept to its variant (e.g. 'nation' to 'nationality'), a renaissance of interest in the transcendent beyond traditional notions of religion and the supernatural ('occultism', 'magic', 'tales'), fascination for the new instead of the old, or a loss of interest in 'royalty' and 'rank'. Toys and concepts like 'tradition', the 'world', 'culture', 'education', 'films', 'games', 'electricity' and 'appliances' make a debut in art. A representation of such tendencies in content change with manifest tensions is visualized in Figure 1. Here, tendency means a projected possible, but not necessarily continuous, trend - should the composition of the collection continue to evolve over the next epoch like it used to develop over the past one, the indicated splits and merges would be more probable to form new content agglomerations than random ones.

5.3 Content Dynamics

As we were left with the impression that in a statistically constructed vector field of term semantics drifts are the norm and not the exception, to account for such dynamics we computed a series of epoch-specific gravitational fields and their generating potential for a first overview. With BMU vector distances between term pairs and their PageRank values for 'term mass', both types of surfaces expressed the interplay between semantic similarity and term importance in a social perspective (Figure 2).

6. CONCLUSIONS AND FUTURE WORK

In the above test, we resolved semantic drift detection, drift measurement, and partly resolved drift interpretation by the automatic evaluation of term cluster consistency. For the detection task, our detailed and thoroughly documented findings indicated that in an evolving collection, as could be expected from the idea of the dynamic library where vector space update results in displaced cluster centroids [36], drifts occur on a regular basis and become more frequent with increasing index term specificity. Apart from surveying the evolving semantic content structure, Somoclu also mapped the parallel evolution of classification tension structure, a precondition to future modeling and anomaly prediction.

(a) 1796-1800 (b) 1801-1805

Figure 1: Excerpt from the tension vs. content structure changes in the level 2 (intermediate) index term landscape in 1796-1805. Blue basins host content, brown ridges indicate tensions. Whereas 'towns', 'cities', 'villages' remain merged over both epochs, 'inland' and 'natural' become merged by 1805.

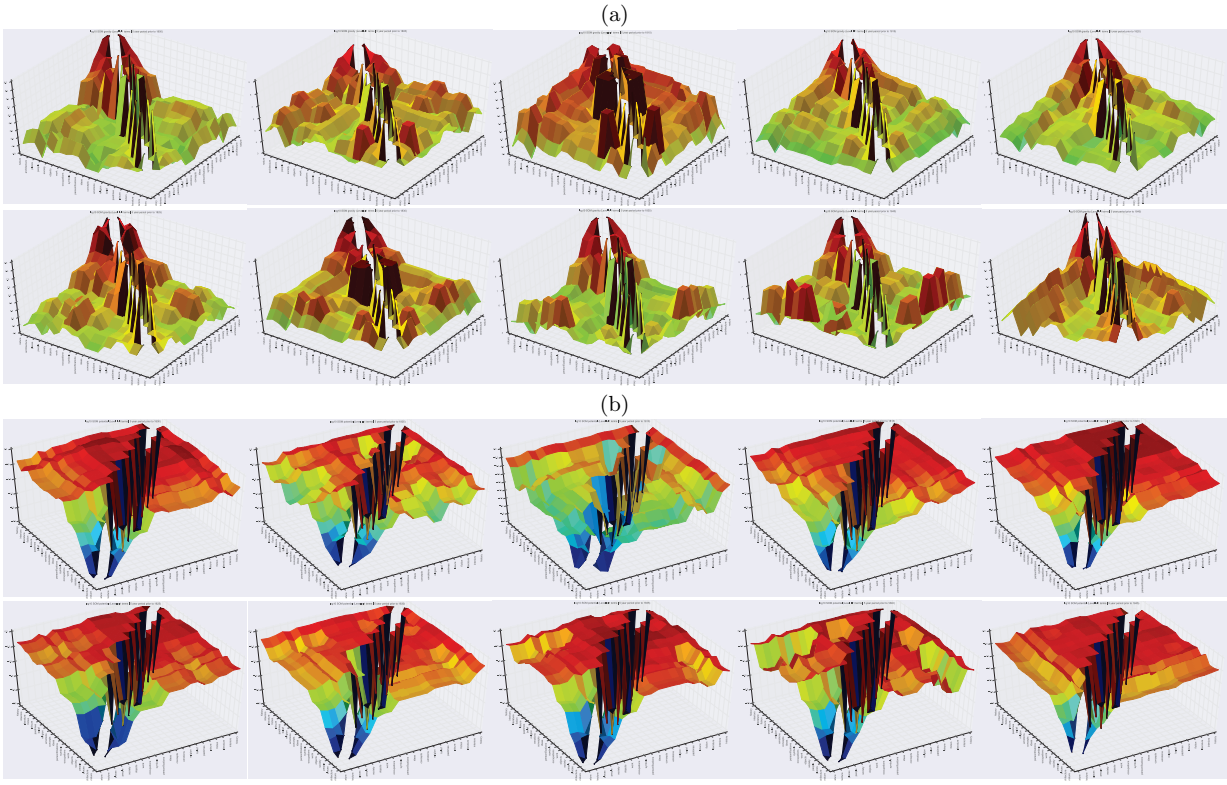

Figure 2: (a) Changes in the top [level 1] conceptual layer of the Tate indexing vocabulary in 1796-1845, sampled every 5 years, modeled on a gravitational field. Gravitational force is the negative gradient of the corresponding potential. (b) Respective changes in the underlying potential field. Extreme values indicate semantically related term pairs with high social status expressed by PageRank.

Further we computed those evolving epoch-specific potential surfaces whose negative gradient was term similarity combined with term importance as an attractive force between feature or object pairs. This potential can be seen as the conceptual consequence of the semantic differential [26], a forerunner to modern latent semantic methods. This semantic potential, in turn, suggests that physics as a metaphor is useful because it yields new, helpful concepts to model the dynamics of meaning, itself important for knowledge organization and knowledge management.

Our effort belongs to the field of *social mechanics*, a 21st century repercussion of ideas dating back as far as 1769 when American political theorist James Madison (1751-1836), the so-called 'father of the constitution' and the United States' fourth president, was said to be studying a primitive form of it at Princeton. After him and over the centuries to come, prominent thinkers often tried to understand society's workings e.g. by means of thermodynamics or mechanics. In our implementation, social mechanics is a variant of classical mechanics because the concept of mass we apply to features in general and index terms in particular, is a relative (evolving) one, depending on language use as its social context and implemented by the distributional hypothesis.

By doing so, the 'meaning as change' paradigm receives experimental support inasmuch as 'term mass' corresponds to work investment during update, with the reconfiguration of semantic spaces and fields being proportional to it. In order to explore the semantic potential, to connect measures of semantic relatedness with centrality values such as PageRank for 'term mass' will be subject to future research, with substantial input expected e.g. from [30] or [9].

7. ACKNOWLEDGMENTS

This research received funding by the European Commission Seventh Framework Programme under Grant Agreement Number FP7-601138 PERICLES. Sándor Darányi is grateful to Emma Tonkin (University of Bristol) for early discussions on the subject.

8. REFERENCES

[1] E. Adar, M. Dontcheva, J. Fogarty, and D. S. Weld. Zoetrope: interacting with the ephemeral web. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*, pages 239–248. ACM, 2008.

[2] A. Aghajanyan. Gravitational clustering. *arXiv preprint arXiv:1509.01659*, 2015.

[3] O. Alonso, J. Strötgen, R. A. Baeza-Yates, and M. Gertz. Temporal information retrieval: Challenges and opportunities. *TWAW*, 11:1–8, 2011.

[4] S.-i. Amari and S. Wu. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6):783–789, 1999.

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

[6] D. Beeferman, A. Berger, and J. Lafferty. A model of lexical attraction and repulsion. In *Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics*, pages 373–380, July 1997.

[7] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117, Apr. 1998.

[8] B. Coecke, M. Sadrzadeh, and S. Clark. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36:345–385, 2010.

[9] D. Cooper. *Linguistic Attractors: The Cognitive Dynamics of Language Acquisition and Change*. Human cognitive processing. J. Benjamins Publishing Company, 1999.

[10] G. V. Cormack, C. L. Clarke, and S. Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759. ACM, 2009.

[11] F. de Jong, H. Rode, and D. Hiemstra. Temporal language models for the disclosure of historical text. In *Humanities, computers and cultural heritage: Proceedings of the XVIth International Conference of the Association for History and Computing (AHC 2005)*, pages 161–168, Amsterdam, The Netherlands, September 2005. Royal Netherlands Academy of Arts and Sciences. Imported from EWI/DB PMS [db-utwente:inpr:0000003683].

[12] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

[13] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. *ACM Transactions on the Web (TWEB)*, 1(2):7, 2007.

[14] J. Eklund. *With or without context: Automatic text categorization using semantic kernels*. PhD thesis, University of Borås, 2016.

[15] I. Frommholz, B. Larsen, B. Piwowarski, M. Lalmas, P. Ingwersen, and K. Van Rijsbergen. Supporting polyrepresentation in a quantum-inspired geometrical retrieval framework. In *Proceedings of the third symposium on Information interaction in context*, pages 115–124. ACM, 2010.

[16] J. A. Gulla, G. Solskinnsbakk, P. Myrseth, V. Haderlein, and O. Cerrato. Semantic drift in ontologies. In *WEBIST (2)*, pages 13–20, 2010.

[17] Z. Harris. *Mathematical structures of language*. Interscience Publishers, New York, NY, USA, 1968.

[18] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

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

[20] M. C. Klein and D. Fensel. Ontology versioning on the semantic web. In *SWWS*, pages 75–91, 2001.

[21] T. Kohonen. *Self-Organizing Maps*. Springer, 2001.

[22] M. Koolen, J. Kamps, and V. de Keijzer. Information retrieval in cultural heritage. *Interdisciplinary Science Reviews*, 34(2-3):268–284, 2009.

[23] J. Lötsch and A. Ultsch. Exploiting the structures of the U-matrix. In *Advances in Self-Organizing Maps and Learning Vector Quantization*, pages 249–257. Springer, 2014.

[24] A. Meroño-Peñuela, C. Guéret, R. Hoekstra, and S. Schlobach. Detecting and reporting extensional concept drift in statistical linked data. In *1st International Workshop on Semantic Statistics (SemStats 2013), ISWC. CEUR*, 2013.

[25] A. Moschitti. Kernel engineering for fast and easy design of natural language applications. In *Proceedings of the 23rd International Conference on Computational Linguistics: Kernel Engineering for Fast and Easy Design of Natural Language Applications*, pages 1–91. Association for Computational Linguistics, 2010.

[26] C. Osgood, G. Suci, and P. Tannenbaum. *The Measurement of Meaning*. University of Illinois Press, Urbana-Champaign, IL, USA, 1957.

[27] P. Pareti, E. Klein, and A. Barker. A linked data scalability challenge: Concept reuse leads to semantic decay. In *Proceedings of the ACM Web Science Conference*, page 7. ACM, 2015.

[28] L. Peng, B. Yang, Y. Chen, and A. Abraham. Data gravitation based classification. *Information Sciences*, 179(6):809–819, 2009.

[29] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43, 2014.

[30] J. Petitot. *Morphogenesis of Meaning*. European semiotics. P. Lang, 2004.

[31] S. Pulman. Distributional semantic models. In C. Heunen, M. Sadrzadeh, and E. Grefenstette, editors, *Quantum Physics and Linguistics: A Compositional, Diagrammatic Discourse*, pages 333–358. Oxford University Press, ISBN 978-0-19-964629-6, 2013.

[32] K. Radinsky, F. Diaz, S. Dumais, M. Shokouhi, A. Dong, and Y. Chang. Temporal web dynamics and its application to information retrieval. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 781–782. ACM, 2013.

[33] S. Robertson and H. Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[34] S. E. Robertson and K. S. Jones. Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3):129–146, 1976.

[35] M. Sadrzadeh and E. Greffenstette. A compositional distributional semantics, two concrete constructions, and some experimental evaluations. In *Proceedings of QI-11, 5th International Quantum Interaction Symposium*, Aberdeen, UK, June 2011.

[36] G. Salton. *Dynamic Information and Library Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1975.

[37] C. Schlieder. Digital heritage: Semantic challenges of long-term preservation. *Semantic Web*, 1(1-2):143–147, 2010.

[38] B. Shaparenko, R. Caruana, J. Gehrke, and T. Joachims. Identifying temporal patterns and key players in document collections. In *Proceedings of the IEEE ICDM Workshop on Temporal Data Mining: Algorithms, Theory and Applications (TDM-05)*, pages 165–174, 2005.

[39] A. Tosi, I. Olier, and A. Vellido. Probability ridges and distortion flows: Visualizing multivariate time series using a variational Bayesian manifold learning method. In *Advances in Self-Organizing Maps and Learning Vector Quantization*, pages 55–64. Springer, 2014.

[40] J. Trier. Das sprachliche Feld. *Neue Jahrbücher für Wissenschaft und Jugendbildung*, 10:428–449, 1934.

[41] P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188, 2010.

[42] M. Tury and M. Bieliková. An approach to detection ontology changes. In *Workshop proceedings of the sixth international conference on Web engineering*, page 14. ACM, 2006.

[43] J. v. Uexküll and G. Kriszat. *Streifzüge durch die Umwelten von Tieren und Menschen. Ein Bilderbuch unsichtbarer Welten: Einundzwanzigster Band*, volume 21. Springer-Verlag, 2013.

[44] A. Ultsch. Clustering with SOM: U* c. In *Proceedings of WSOM-05, 5th Workshop on Self-Organizing Maps*, pages 75–82, Paris, France, September 2005.

[45] M. Uschold. Creating, integrating and maintaining local and global ontologies. In *Proceedings of the First Workshop on Ontology Learning (OL-2000) in conjunction with the 14th European Conference on Artificial Intelligence (ECAI-2000)*. Citeseer, 2000.

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

[47] S. Wang, S. Schlobach, and M. Klein. Concept drift and how to identify it. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(3):247–265, 2011.

[48] H. White. Cross-textual cohesion and coherence. In *Proceedings of the Workshop on Discourse Architectures: The Design and Analysis of Computer-Mediated Conversation*, Minneapolis, MN, USA, April 2002.

[49] P. Wittek, S. Darányi, E. Kontopoulos, T. Moysiadis, and I. Kompatsiaris. Monitoring term drift based on semantic consistency in an evolving vector field. In *Proceedings of IJCNN-15, International Joint Conference on Neural Networks*, 2015.

[50] P. Wittek, S. Darányi, and Y. Lin. A vector field approach to lexical semantics. In *Proceedings of QI-14, 8th International Conference on Quantum Interaction*, pages 78–92, June 2014.

[51] P. Wittek, S. C. Gao, I. S. Lim, and L. Zhao. Somoclu: An efficient parallel library for self-organizing maps. *arXiv:1305.1422*, 2015.

[52] P. Wittek, B. Koopman, G. Zuccon, and S. Darányi. Combining word semantics within complex Hilbert space for information retrieval. In *Proceedings of QI-13, 7th International Quantum Interaction Symposium*, pages 160–171, July 2013.

[53] L. Wittgenstein. *Philosophical Investigations*. Blackwell Publishing, Oxford, UK, 1967.

[54] B. Yildiz. Ontology evolution and versioning. *Vienna University of Technology, Karlsplatz*, 2006.

[55] D. Yuret. Discovery of linguistic relations using lexical attraction. *arXiv:cmp-lg/9805009*, 1998.