Reading Through Graphics: Interactive Landscapes to Explore Dynamic Topic Spaces

Eva Ulbrich[1], Eduardo Veas[1](✉), Santokh Singh[1], and Vedran Sabol[1,2]

[1] Know Center GmbH, Inffeldgasse 13, 8010 Graz, Austria
{eulbrich,eveas,ssingh,vsabol}@know-center.com
[2] University of Technology Graz, Graz, Austria

Abstract. An information landscape is commonly used to represent relatedness in large, high-dimensional datasets, such as text document collections. In this paper we present interactive metaphors, inspired in map reading and visual transitions, that enhance the landscape representation for the analysis of topical changes in dynamic text repositories. The goal of interactive visualizations is to elicit insight, to allow users to visually formulate hypotheses about the underlying data and to prove them. We present a user study that investigates how users can elicit information about topics in a large document set. Our study concentrated on building and testing hypotheses using the map reading metaphors. The results show that people indeed relate topics in the document set from spatial relationships shown in the landscape, and capture the changes to topics aided by map reading metaphors.

Keywords: Text visualisation · Dynamic information landscape · Interaction design · User study

1 Introduction

The already enormous amount of electronically available information keeps growing at ever faster rates. While retrieval tools excel at finding a single or a few relevant pieces of information, when a holistic view on large amount of complex data is needed, it becomes necessary to consider the entirety of the data set for analysis. Information Landscapes represent a powerful visualization technique for interactive analysis of complex topical relationships in large document repositories [7]. They convey topical similarity in a document set through spatial proximity [4,12]. However, we are not just confronted with large, but also with permanently growing and rapidly changing repositories. The concept of information landscapes has been adapted to addresses the visualisation of changes in the topical structure of a data set within a single, consistent visual metaphor - the dynamic topography information landscape [16]. Adding new documents to the set, shifts the position of topics in the dynamic landscape. But, representing temporal evolution in a comprehensive manner, although crucial for the analysis of large corpora, is not trivial. State-of-the-art research in the area identified caveats and pitfalls of representing change with information landscapes, particularly related to change blindness [11].

S. Yamamoto (Ed.): HIMI 2015, Part I, LNCS 9172, pp. 127–137, 2015.
DOI: 10.1007/978-3-319-20612-7_13

This paper presents an efficient web-based implementation of the dynamic information landscape. We introduce novel map-reading interactive metaphors for topic analysis including interactive morphological transitions, multiple views, trails and traces. In addition, a major contribution is a user study conducted with 18 test users to validate the map reading metaphors as analysis tools for information landscapes. In particular, we follow two goals: (i) to find out whether people can read information about topics and topical relationships out of spatial relationships shown in the landscape, and (ii) to discover if people can derive information about the evolution of topics in dynamically changing document collections. Our evaluations showed a relation between observations and assumptions elicited with aforementioned metaphors: participants built assumptions about the data that they could validate. Map reading metaphors, as applied here, are thus useful for topic analysis of a large body of documents. They give an overview of the distribution of concepts and topics. Transition techniques complemented with map reading metaphors were useful tools to analyse the evolution of topics, by visually comparing where new topics appeared, which topics moved closer together, and what new topics became important.

2 Related Work

The fundamental idea of information landscapes is to convey relatedness between data elements, in our case the topical similarity, through spatial proximity in the visualisation. The notion builds on the so-called "first law of cognitive geography", stating that people assume that close things are similar, validated in [9]. Reference [14] describes the research agenda for spatialisation methods, in particular the principles of the geographic metaphor for visualization of non-geographic information. Reference [5] describes a formalisation of spatialisation views and the theoretical foundations of the discipline. An interesting result is a discussion on visual features to better represent properties of the data, such as conveying similarity with spatial proximity, magnitude with height, and change with tectonic processes.

Information landscapes have been used to visualise topical distribution of document collections for more than 20 years, starting with Bead [4] in 1993 and SPIRE [17] in 1995. Using the VxInsight tool they have been successfully applied to the analysis of patent data bases [2] and of scientific and technological document sets [3]. InfoSky [1] demonstrated the applicability of the information landscape concept on hierarchically organised document collections, where the hierarchy is represented by nested Voronoi polygons.

The temporal behaviour in document sets is often represented by dedicated visualizations, such as the well-known ThemeRiver [6], which successfully conveys topical trends and correlations. To enable interactive visual analysis of both topical relationships and topical trends, an information landscape was combined with a ThemeRiver using a Multiple Coordinated Views interface [13]. The concept of a dynamic topography information landscape for visualisation of changes in document repositories was proposed in [12] and realised in [13] and [16].

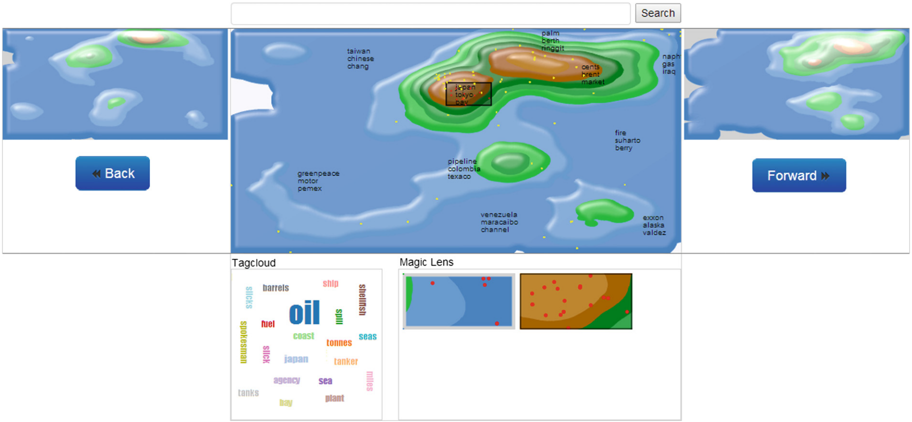

Fig. 1. The test UI: Multiple views metaphor shows previous and subsequent stages of the landscape. Back and Forward buttons will trigger the morphing process. A magic lense is used for selection whereby different selections can be saved (shown as thubmnails) and reapplied. Keywords from selected documents are shown in the tagcloud. Clicking on keywords triggers filtering of documents in the list (not shown).

These early versions lacked all but most basic interaction techniques, suffered from issues related to changes blindness and, most notably, their dynamic features were not evaluated in a user study. Issues related to change blindness in dynamic information landscapes were discussed in [11] with remedies proposed, but evaluated with a rather limited example, which solely changes height from appearing or disappearing documents. Documents positions were fixed, hence changes in document positions occurring due to the altered topical structure are not considered.

3 Dynamic Information Landscape

An information landscape conveys similarity of topics in a data set through spatial proximity and density of topics as elevation in the visualization. Hills represent groups (clusters) of topically related documents separated by areas represented as sea. Higher hills group more documents than lower ones. Landscape areas are labeled with descriptive terms extracted from the documents.

Besides topical relatedness, dynamic information landscapes represent changes to a document repository as changes in topography. As a repository evolves (e.g. documents are added) the collection of topics changes, new topics become important, documents move, attracted by new topical relationships. Hence the landscape topography is altered. A fading topic may cause islands and hills to disappear, while new hot topics cause new islands to arise from the seabed. Hills moving towards or apart from each other indicate topical convergence or divergence of the corresponding topical clusters, with merging (cluster fusion) and splitting (cluster break-up) of hills and islands occurring in extreme cases.

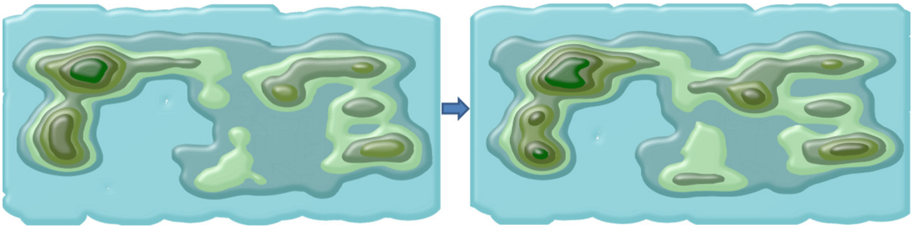

Fig. 2. Morphological changes after adding new documents.

To be comprehensible, transitions of the landscape topography from an old to a new temporal configuration must be incremental. The configuration of regions which are not (or are only little) affected by the modification of the data set must remain stable with respect to their relative positions and shapes. This enables users to immediately understand the altered landscape through the recognition and orientation provided by the already known, preserved (or scarcely modified) elements of the topography. A morphing procedure modifies the topography in smoothly animated transitions, that users can follow to understand the changes.

3.1 Computation Procedure

To compute the 2D similarity layout for the documents we use a fast, scalable, aggregation-based projection algorithm [10] which employs k-means clustering and cluster-oriented force-directed placement. The algorithm was modified to support incremental computation enabling seamless incorporation of changes into an existing similarity layout [15]. The landscape generation begins with the vectorisation of documents, and includes the above mentioned incremental clustering and projection, cluster labelling (using highest weight terms from the underlying documents), height matrix computation, and finally the extraction of contour lines (isohypses). As this process is computationally intensive it is performed on the server. The result of the computation is a JSON file defining the landscape geometry. The geometry is transferred to and displayed by the Web-Client built using the D3.js JavaScript library (http://d3js.org/). For pairs of consecutive landscapes we compute mappings between the corresponding contour lines and use morphing to smoothly visualise the transitions.

3.2 Map Reading Metaphors

Figuring out relations between documents in a large corpus is a cognitively demanding task. The algorithms proposed can establish a level of similarity between entities and documents. Metaphors based on map reading were introduced to foster the exploratory analytics workflow. The goals are to enable visual thinking in the exploratory phase, whereupon hypotheses are formulated, and to promote the explanatory role thereafter, whereby hypotheses are supported or revoked based on careful observation of data characteristics. Therefore metaphors for *map navigation*, *overview and details* and *temporal transitions* were introduced (Fig. 2).

Map Navigation and Selection. Map navigation can be broken down to a series of pan and and zoom operations. Special attention was put to enable incremental changes in information density with changing zoom levels, organizing labels and isohypses in respective level of detail hierarchies. *Search for terms* works as a keyword search on labels and feature vectors. It highlights documents and areas where the keyword occurs. *Area selection* highlights document points in a region. Selections can be stored to recover information after transitions.

Overview and Details. A multiple views metaphor was introduced to reveal details for areas of interest while maintaining overview (see Fig. 1). It consisted of a magic lens used for inspection and selection. The magic lens triggered miniature (zoomed) previews and a wordcloud view summarizing content. Selections can be stored to in a rectangular thumbnail to create a trail of selections of interest to allow tracking a chain of changes.

Temporal Transitions. The multiple views metaphor extends to temporal transitions. Mini landscapes show abstract views of preceding and subsequent stages in the progression. Upon selection and when using the magic lens, previews and wordclouds are added with for past and current temporal stages. Additionally, a *document trail* metaphor shows document flow between time stages. Additionally, new documents are highlighted in the landscape upon transition, for quick overview of the changes in density (see Fig. 3). The last two metaphors were particularly useful while fine-tuning the incremental layout algorithm: the developer could actually observe what similarities were enforced by it and adjust accordingly.

4 Evaluation

The motivation to enable interactive analysis on dynamic information landscapes stresses the analytic evolution of concepts and topics over time. A formative evaluation was carried out to establish that aided by map reading tools:

H1 people can elicit information about topics in documents out of spatial relationships shown in the landscape,

H2 people can derive information about the evolution of topics in a document set.

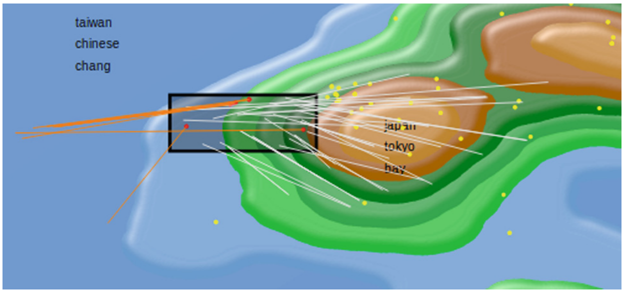

Fig. 3. Document trails show document movement between incremental stages.

The evaluation concentrates on the exploratory analytics workflow. It enforces the role of metaphors in an exploratory phase, whereupon facts are observed to formulate hypotheses. It then exploits tools for the explanatory phase of validating or discarding hypotheses.

4.1 Participants and Methodology

Eighteen participants took part in the experiment (14M, 4F, $X = 29$ years). Nine of them had corrected vision. Sixteen participants had experience in data analysis, fourteen used visual tools.

Following the aforementioned exploratory analytics workflow, the evaluation had iterative stages of exploration, hypotheses formulation, and hypotheses validation. To investigate the complexity of the hypotheses, we grouped them into:

Structural: only dealing with changes to geometry, without any association to topics.

N-entity: mentioning at most n entities.

Further, we analyzed the content of n-entity hypotheses to find out what kind of relations participants elicit between recovered entities.

The study consisted of a training phase and the proper evaluation. The training phase introduced the metaphors with a screen cast followed by hands-on practice. Thereafter, participants could freely use the tool until they felt confident with it. It took an average 15 min.

The proper evaluation consisted of three analytics phases based on three landscapes which were incremental progressions (see Sect. 4.2). In the exploration stage participants used tools to discover content. Participants were thus asked to identify areas, the number of documents in an area, a document with specific content, and an area with certain attributes. In the hypotheses formulation stage, participants were required to formulate 3 hypotheses based on changes introduced by adding new documents, by visually comparing the current landscape with the thumbnail of the subsequent one. The hypotheses validation stage followed by transitioning to the subsequent landscape. At this stage people used tools to validate their hypotheses.

The study closed with a subjective questionnaire whereby participants rated the usefulness on a 7-point-Likert scale and suggested improvements. The complete study took in average 52 min.

4.2 Stimuli and Apparatus

The stimuli were created from a subset of the Reuters Corpus Volume 1 documents, whereby an initial landscape was computed. Thereafter, incremental document sets were added in a controlled manner to create a sequence of landscapes. Hereby, Six incremental landscapes were obtained. The first three were used for the proper evaluation.

To obtain the training stimuli, the last two landscapes were relabelled with Christmas terms, to prevent familiarity with the concepts in the proper evaluation. Landscape 4 was left out to achieve aesthetic separation between training and evaluation stimuli. Additionally, the "sealevel" of the training data was placed a level higher, further altering the visual aspect of the landscapes.

The test was conducted in a calm, small room with a conventional screen connected to a notebook. Participants used mouse and keyboard to complete tasks. The visualization was running on Chrome Version 30.0.1599.101.

4.3 Results

Exploration Phase. Participants could identify regions of high density of documents by searching for high elevations. Thirteen participants found the seven high elevation regions (7/7), three miscounted and found eight (8/7), only two had trouble and counted more (11/7) or less (3/7). Participants identified real countries in labels. Ten of them found all six countries (6/6), six did not count Taiwan and found five (5/6), one counted four (4/6) and one counted eight (8/6). In the next task, participants had to find high density area with the terms ship and spill. All participants completed this task using search tools. When asked to identify important terms in a selected region, participants used the word cloud and identified Oil, Japan and tanker. One participant opted for using just labels (and not the word cloud) and chose Japan, Slicks, Reactors. Participants also had to identify topic mountains from abstract descriptions, namely they had to find a region about the oil price and one about a ship accident where neither oil price nor ship accident appear as labels. All participants chose a topic mountain solely containing information about money, market and cents to be the one about oil price and sixteen (16/18) chose the mountain with the labels Japan, Slicks, Reactors for the accident. One chose a region around the Exxon Valdez labels as he remembered it to be a tanker accident, another one chose a region around the label Greenpeace as he thought they might had caused a ship to sink. The latter two made their choice solely by interpreting labels, whereas the others used the search tool to find terms within documents related to accidents with ships, or selected areas and checked for significant terms in the word cloud.

Hypothesis Formulation. In this stage, participants had to visually compare the current landscape with the subsequent one and build three hypotheses. This lead to a total of 54 hypotheses in the first transition and 53 in the second (one participant only formulated two). Our intention here was to find out if participants could analyse the evolution of topics in a document set with the provided metaphors (Table 1).

In the first transition, we found 7 (0.13) purely structural hypotheses. Three of them from a single user who did not make any association of geometry to topics. 19 hypotheses (0.35) mention a single entity (1-entity), mostly observations about an increase or decrease of documents. 21 hypotheses (0.39) relate two entities (2-entity) in more complex relations. Finally we found 6 (0.11) 3-entity and 1 (0.02) 4-entity. Regarding the contents of hypotheses, 17 (0.31) and 9 (0.17) hypotheses predicted an increase or decrease of documents respectively in relation to a topic. 14 hypotheses predicted relocations of topics

Table 1. Classification of Hypotheses. Structural hypotheses only referred to changes in geometry, n-entity hypotheses established relations between n entities. For both transitions participants tended to build more hypotheses relating 2-entities.

Entity	Count (%)	Count (%)
	Transition 1	Transition 2
Structural	7 (0.13)	3 (0.05)
1-Entity	19 (0.35)	8 (0.15)
2-Entity	21 (0.39)	23 (0.435)
3-Entity	6 (0.11)	13 (0.245)
4-Entity	1 (0.02)	5 (0.10)
5-Entity	0	1 (0.02)
Total	54	53

and 0.037 the appearance of new topics. Interestingly, 14 hypotheses stated complex relationships between entities, including 3 (0.055) similarity and 11 (0.203) varied complex ones, such as changes in oil-price due to accidents of tankers or delays with pipelines.

In the second transition, we found 3 (0.05) purely structural hypotheses all from the same user. 8 hypotheses (0.15) mention a single entity (1-entity). We also found 23 (0.43) 2-entity, 13 (0.24) 3-entity, 5 (0.1) 4-entity, and 1 (0.02) 5-entity hypotheses. Regarding contents, 19 (0.35) hypotheses predicted an increase in documents about a topic. Curiously no hypothesis mentioned decrease in a topic. 3 hypotheses predicted relocations of topics. Interestingly 29 hypotheses (0.54) reported complex relationships, including 10 (0.188) similarity and 19 (0.358) more complex assumptions (Table 2).

Hypothesis Validation. Building hypothesis was only a half of this test, after transitioning to the target landscape, participants had to validate or discard their hypotheses using the proposed tools. Results showed that people could actually analyse the evidence to (dis) prove hypotheses. In the first transition, 6 hypotheses could not be proved or rejected (score 3-4), 26 were proved (scored < 3)

Table 2. Sample Hypotheses. Hypotheses were also classified regarding their contents, in whether they related an increase or decrease of importance of a topic or more complex assumptions.

Entity	Type	Hypothesis
structural	–	many new documents on one hill, rest remains similar
1-entity	increase	more news about Iraq
2-entity	decrease	Colombia and its pipeline lose importance as mountain shrinks
2-entity	assumption	a new development related to reactors in Japan
4-entity	similarity	more news relating Irak, Japan and Venezuela with gas

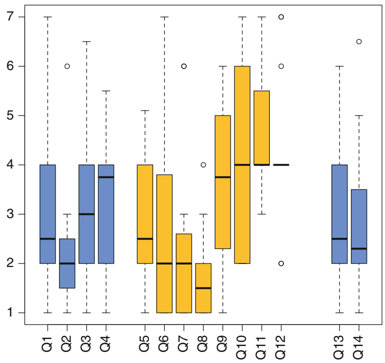

Fig. 4. Subjective ratings. Plots in blue rated overall usability while plots in yellow rated the metaphors. Refer to the text in Exit Questionnaire, Sect. 4.3

and 22 were disproved (scored > 4). In the second transition, 12 hypotheses could not be proved or rejected, 16 were proved and 24 were rejected. This shows that participants found evidence using the tools to either back or revoke the majority of hypotheses (Fig. 4).

Exit Questionnaire. Participants found that they could build useful hypothesis (Q1, Median = 2.5). They also found the landscape gives a solid overview about distribution of topics (Q2, M = 2), and that the thumbnail landscape helped to build hypotheses (Q3,M = 3). Participants were neutral about how exhausting it was to build hypotheses (Q4, M = 3.75). With regards to the tools used to prove hypotheses and how helpful they were, search for terms was helpful (Q5, M = 2.5), as well as the Word Cloud (Q6, M = 2) and the document flow (Q7, M = 2). The rectangle selection rated very well (Q8, M = 1.5). The magic lens was partially useful (Q9, M = 3.4), while the miniature preview borderline (Q10, M = 4), and saving labels positions was rated rather unuseful (Q11, M = 6). The list also scored poorly (Q12, M = 4). Finally, participants found that they could get information out of the data without having to read the documents themselves (Q13, M = 2.5). They also found the visualisation as a whole a useful tool to analyze data (Q14, M = 2.2).

4.4 Discussion

We can report that participants in general can relate changes in geometry to topical changes, as they formulated hypotheses majorly relating entities out of spatial relations in the visualization. Furthermore, participants formulated hypotheses relating two and up to four entities in complex relations. Although hypotheses were often invalid, the study showed that our metaphors aided participants to analyse the evidence, judge and retain or reject hypotheses. The fact that people could not decide about some hypotheses is not a limitation. These are the cases where they would have to find more evidence, e.g. by reading some of the documents. The proposed metaphors actually empower users to discriminate for which assumptions or hypotheses they would need more information.

The metaphors further direct users to the sources of that information, since the documents are actually linked and can be accessed from the tool.

The exit questionnaire validates our findings: participants found in general they could build useful hypotheses, although it was not trivial. In general participants found the interactive metaphors of dynamic information landscape are useful to analyze and obtain information from a large body of documents without having to read them all.

5 Conclusions

We built an HTML5-based information landscape using data from a text processing pipeline. The pipeline starts by gathering documents and clustering based on their content similarities. It then creates a three dimensional height matrix. Topic mountains are extracted by cutting these into isohypses.

Based on map analysis, a number of topic analysis metaphors were developed (e.g., focusing on a region to discover concepts, associating high density of topics/documents with elevation). Furthermore, we put special focus on interactive aspects of the map reading metaphors, to compare and obtain information from topic landscapes created incrementally. Our evaluations showed a correlation across observations and assumptions elicited with aforementioned metaphors: participants built assumptions about the data that they could validate. Map reading metaphors, as applied here, are thus useful for topic analysis of a large body of documents. They give an overview of the distribution of concepts and topics. Transition techniques complemented with map reading metaphors were useful tools to analyse the evolution of topics, by visually comparing where new documents appeared, where documents wandered to, and what new topics became important.

In the future, we plan to investigate the impact of individual tools. Additionally, we are currently investigating methods to tightly integrate the interactive metaphors with the algorithmic analytics methods.

Acknowledgement. This work was performed within the FIT-IT funded DIVINE and the EC FP7 EEXCESS (grant 600601) projects. The Know-Center is funded within the Austrian COMET Program – managed by the Austrian Research Promotion Agency (FFG) – under the auspices of BMVIT and BMWFW Austrian Federal Ministries and the state of Styria.

References

1. Andrews, K., Kienreich, W., Sabol, V., Becker, J., Kappe, F., Droschl, G., Granitzer, M., Auer, P., Tochtermann, K.: The infosky visual explorer: exploiting hierarchical structture and document similarities. J. Inf. Visulization **1**(3/4), 166–181 (2002). London, England
2. Boyack, K.W., Wylie, B.N., Davidson, G.S., Johnson, D.K.: Analysis of patent databases using VxInsight. In: Workshop on New Paradigms in Information Visualization and Manipulation (2000)

3. Boyack, K.W., Wylie, B.N., Davidson, G.S.: Domain visualization using VxInsight for science and technology management. J. Am. Soc. Inform. Sci. Technol. **53**, 764–774 (2002)
4. Chalmers, M.: Using a landscape metaphor to represent a corpus of documents. In: Campari, I., Frank, A.U. (eds.) COSIT 1993. LNCS, vol. 716, pp. 377–390. Springer, Heidelberg (1993)
5. Fabrikant, S.I., Buttenfield, B.P.: Formalizing semantic spaces for information access. Ann. Assoc. Am. Geogr. **91**(2), 263–280 (2001)
6. Havre, S., Hetzler, B., Nowell, L.: ThemeRiver: visualizing theme changes over time. In: Proceedings of the IEEE Symposium on Information Visualization 2000 (InfoVis 2000), pp. 115–123 (2000)
7. Krishnan, M., Bohn, S., Cowley, W., Crow, V., Nieplocha, J.: Scalable visual analytics of massive textual datasets. In: 21st IEEE International Parallel and Distributed Processing Symposium. Long Beach, USA, pp. 1–10 (2007)
8. Kroell, M.,Sabol, V., Kern, R., Granitzer, M.: Integrating user preferences into distance metrics. In: Proceedings of the LWA 2013 Workshop on Knowledge Discovery, Data Mining and Machine Learning (2013)
9. Montello, D.R., Fabrikant, S.I., Ruocco, M., Middleton, R.S.: Testing the first law of cognitive geography on point-display spatializations. In: Kuhn, W., Worboys, M.F., Timpf, S. (eds.) COSIT 2003. LNCS, vol. 2825, pp. 316–331. Springer, Heidelberg (2003)
10. Muhr, M., Sabol, V., Granitzer, M.: Scalable recursive top-down hierarchical clustering approach with implicit model selection for textual data sets. In: Proceedings of the 2010 Workshop on Database and Expert Systems Applications (held at DEXA 2010), pp. 15–19 (2010)
11. Nowell, L., Hetzler, E., Tanasse, T.: Change blindness in information visualization: a case study. In: Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS 2001), pp. 15–22 (2001)
12. Sabol, V., Syed, K.A.A., Scharl, A., Muhr, M., Hubmann-Haidvogel, A.: Incremental computation of information landscapes for dynamic web interfaces. In: Proceedings of the 10th Brazilian Symposium on Human Factors in Computer Systems, pp. 205–208 (2010)
13. Sabol, V.: Visual analysis of relatedness and dynamics in complex, enterprise-scale repositories. Doctoral Dissertation, Graz University of Technology, May 2012
14. Skupin, A., Fabrikant, S.I.: Spatialization methods: a cartographic research agenda for non-geographic information visualization. Cartography Geogr. Inf. Sci. **30**(2), 99–119 (2003)
15. Syed, K.A.A., Kröll, M., Sabol, V., Gindl, S., Scharl, A.: Incremental and scalable computation of dynamic topography information landscapes. J. Multimedia Process. Technol. 3(1), Special Issue on the Theory and Application of Visual Analytics, 49–65 (2012)
16. Syed, K.A.A., Kröll, M., Sabol, V., Scharl, A., Gindl, S., Granitzer, M., Weichselbraun, A.: Dynamic topography information landscapes – an incremental approach to visual knowledge discovery. In: Cuzzocrea, A., Dayal, U. (eds.) DaWaK 2012. LNCS, vol. 7448, pp. 352–363. Springer, Heidelberg (2012)
17. Wise, J.A., Thomas, J.J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., Crow, V.: Visualizing the non-visual: spatial analysis and interaction with information from text documents. In: Proceedings of the 1995 IEEE Symposium on Information Visualization, pp. 51–58 (1995)