InfoSky: Visual Exploration of Large Hierarchical Document Repositories

Frank Kappe[1], *Georg Droschl*[1], *Wolfgang Kienreich*[1] *Vedran Sabol*[2], *Jutta Becker*[2], *Keith Andrews*[3], *Michael Granitzer*[2], *Klaus Tochtermann*[2], and *Peter Auer*[3]

Abstract

InfoSky is a system enabling users to explore large, hierarchically structured document collections. Similar to a real-world telescope, InfoSky employs a planar graphical representation with variable magnification. Documents are assumed to have significant textual content, which can be extracted if necessary with specialised tools. Documents of similar content are placed close to each other and are visualised as stars, forming clusters with distinct shapes. For greater performance, the hierarchical structure is exploited and force-directed placement is applied recursively at each level on much fewer objects, rather than on the whole corpus.

1 InfoSky

InfoSky, enables users to explore large, hierarchically structured document collections. Similar to a real-world telescope, InfoSky employs a planar graphical representation with variable magnification and the metaphor of a zooming galaxy of stars, organised hierarchically into clusters. Documents of similar content are placed close to each other and are visualised as stars, forming clusters featuring distinct shapes, which are easy to recall.

InfoSky assumes that documents are already organised in a hierarchy of collections and sub-collections, called the *collection hierarchy*. Both documents and collections can be members of more than one parent collection, but cycles are explicitly disallowed. This structure is otherwise known as a directed acyclic graph. The collection hierarchy might, for example, be a classification scheme or taxonomy, manually maintained by editorial staff. The collection hierarchy could also be created or generated (semi-)automatically. Documents are assumed to have significant textual content, which can be extracted if necessary with specialised tools. Documents are typically plain text, PDF, HTML, or Word documents, but may also include spreadsheets and many other formats.

InfoSky combines both a traditional tree browser and a new telescope view of a galaxy, as shown in Figure 1. In the galaxy, documents are visualised as stars and similar documents form clusters of stars. Collections are visualised as polygons bounding clusters and stars, resembling the boundaries of constellations in the night sky. Collections featuring similar content are placed close to each other, as far as the hierarchical structure allows. Empty areas remain where documents are hidden due to access right restrictions, and resemble dark nebulae found quite frequently within real galaxies. The telescope is used as a metaphor for interaction with the visualisation. Users can pan the view point within the visualised galaxy, like an astronomer can point a telescope at any point of the sky.

[1]Hyperwave R&D, Albrechtgasse 9, A-8010 Graz, Austria, {fkappe|gdroschl|wkien}@hyperwave.com

[2]Know-Center, Inffeldgasse 16c, A-8010 Graz, Austria, {vsabol|jbecker|mgrani|ktochter}@know-center.at

[3]Graz University of Technology, Inffeldgasse 16c, A-8010 Graz, Austria, kandrews@iicm.edu and pauer@igi.tu-graz.ac.at

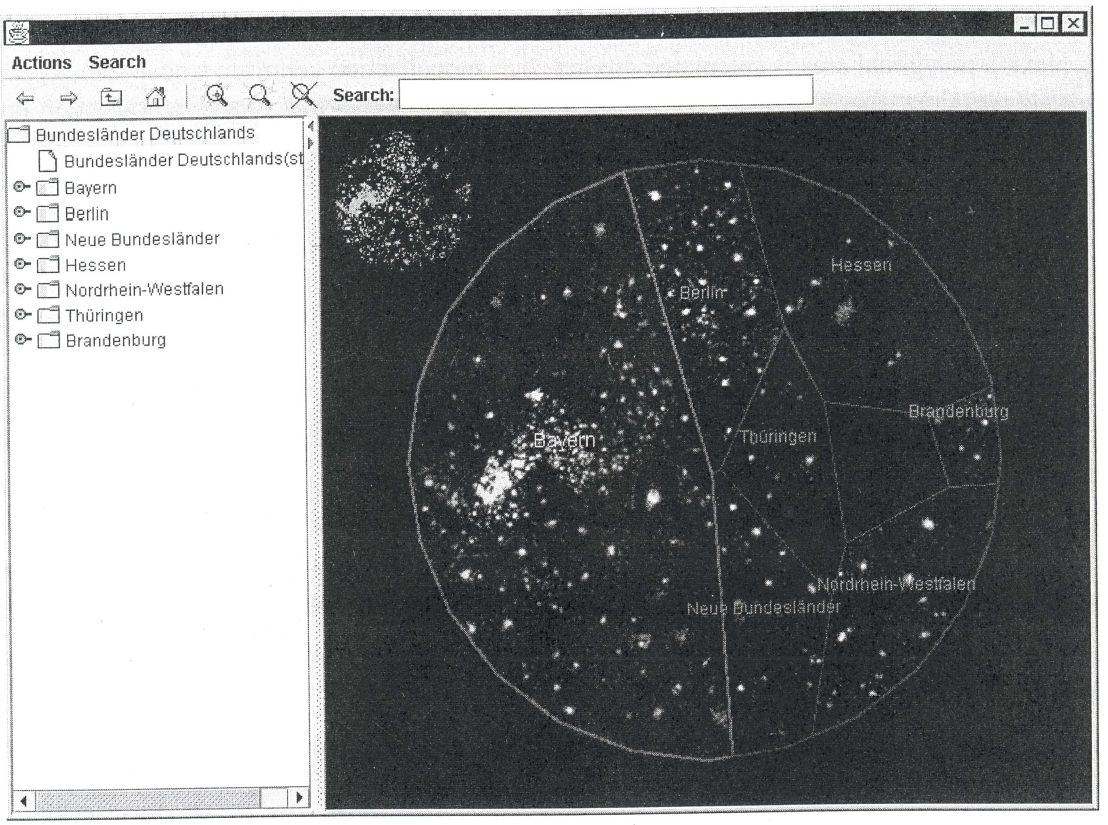

Figure 1: The original prototype of the InfoSky explorer, as used in the comparative study.

Magnification can be increased to reveal details of clusters and stars, or reduced to display the galaxy as a whole.

2 InfoSky Implementation

InfoSky is implemented as a client-server system in Java. On the server side, galaxy geometry is created and stored for a particular hierarchically structured document corpus. On the client side, the subset of the galaxy visible to a particular user is visualised and made explorable to the user.

The galactic geometry is generated from the underlying repository recursively from top to bottom in several steps:

1. First, at each level, the centroids of any subcollections are positioned in a normalised 2D plane according to their similarity with each other using a force-directed similarity placement algorithm. The similarities to their parent's sibling collection centroids are used as static influence factors to ensure that similar neighbouring subcollections across collection boundaries tend towards each other (they are not allowed to actually cross the boundary). The centroid of a synthetic subcollection called "Stars", which holds the documents at that level of the hierarchy, is also positioned.

2. The layout in normalised 2D space is transformed to the polygonal area of the parent collection

using a simple geometric transformation.

3. Next, a polygonal area is calculated around each subcollection centroid, whose size is related to the total number of documents and collections contained in that subcollection (at all lower levels). This polygonal partition of the parent collection's area is accomplished using modified, weighted Voronoi diagrams (Okabe, Boots, Sugihara, & Chiu, 2000, pg. 128), resulting in a recursive spider's web like subdivision of each area.

4. Finally, documents contained in the collection at this level are positioned within their parent's area using the similarity placement algorithm as points within the synthetic "Stars" collection, according to their inter-document similarity and their similarity to the subcollection centroids at this level, which are used as static influence factors.

Basing the layout on the underlying hierarchical structure of the repository has a major advantage in terms of performance. Similarity placement typically has a run-time complexity approaching $O(n^2)$, where n is the number of objects being positioned. However, since similarity placement is only used on one level of the hierarchy at a time, the value of n is generally quite small (the number of subcollection centroids plus the number of documents at that level). Full details of these algorithms are described in (Andrews et al., 2002).

3 User Testing

A small formal experiment with 8 users in a counterbalanced design was run to establish a baseline comparison between the InfoSky telescope browser and the InfoSky tree browser. Users were only allowed to use one or the other in isolation. The dataset used consisted of approximately 100,000 German language news articles provided by the Süddeutsche Zeitung. The articles are manually classified thematically by the newspaper's editorial staff into around 9,000 collections and subcollections upto 15 levels deep. Two sets of tasks were formulated (five pairs of equivalent tasks). The tasks were designed to be equivalent between the two sets in the sense that their solutions lay at the same level of the hierarchy and involved inspecting approximately the same number of choices at each level.

On average, the tree browser performed better than the prototype telescope browser for each of the tasks tested. The overall difference between tree browser and telescope browser was significant at $p < 0.05$ (paired samples t-test, 39 degrees of freedom, $t = 3.038$). The reasons for the slower performance of the telescope browser appear to be two-fold. Firstly, users typically have spent many, many hours using a traditional explorer-like tree browser and are very familiar with its metaphor and controls. Secondly, whereas the tree view component has already undergone many iterations of development, the telescope browser is a prototype at a fairly early stage of development.

When interviewed after the test, users indicated that they were very familiar with a tree browser and liked being able to use the mouse cursor as a visual aid when scanning lists. They liked the overview which the telescope browser provided and could imagine using it for exploring a corpus of documents. This study did not include a task asking users to find similar or related documents or subcollections, something which the telescope metaphor should support quite well. Users further indicated that a combination of both browsers and search functionality could be very powerful. This is something which InfoSky provides, but was not tested in this study.

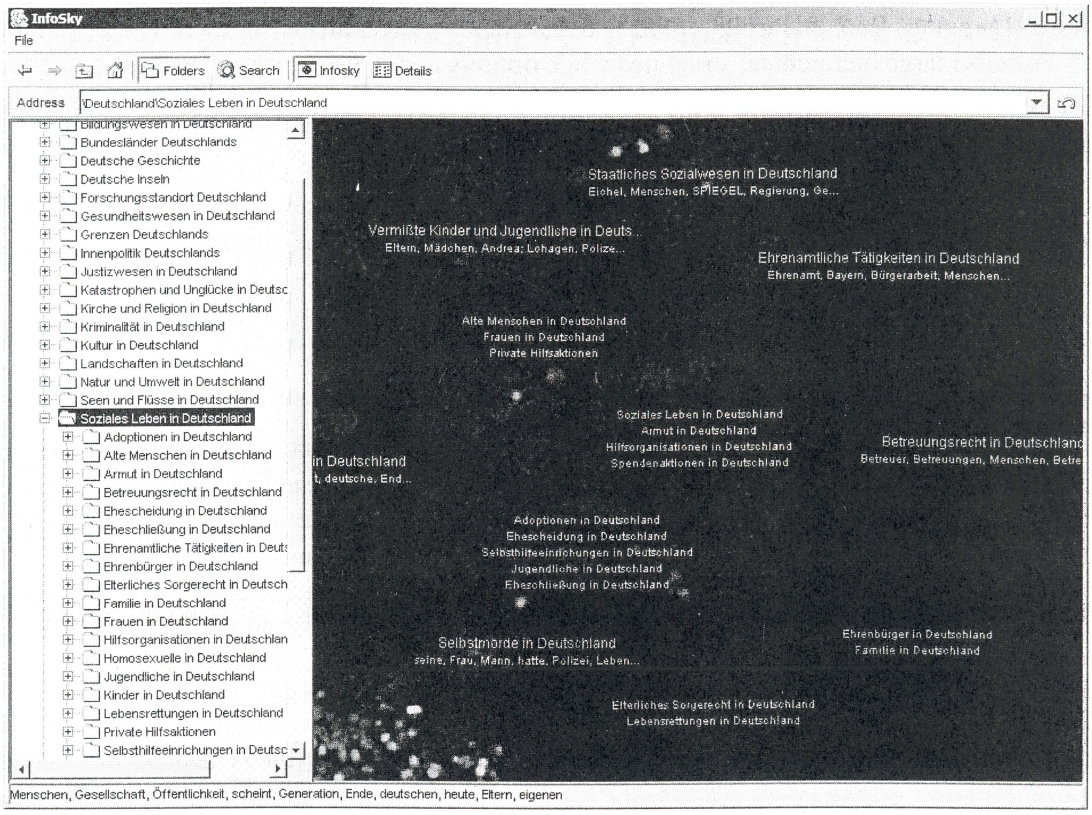

Figure 2: The revised version of InfoSky.

During the study, users complained that some of the Voronoi polygons were too small to see properly. Users were confused by the "Stars" collection used to hold documents at each level. When trying to access individual documents, users were also disturbed by the fact that the document labels appeared to jump around arbitrarily. In fact the prototype displayed only the titles of documents "near" to the cursor, but users were moving the cursor to step through what they perceived to be a list of documents. These problems were addressed in the new version of InfoSky shown in Figure 2.

We have not yet tested the complete InfoSky armoury of synchronised tree browser, telescope browser, and search in context against other methods of exploring large hierarchical document collections. Nor have we tested tasks involving finding related or similar documents or subcollections, something the telescope metaphor should be well-suited to. As development proceeds, we believe that the InfoSky prototype will constitute a step towards practical, user-oriented, visual exploration of large, hierarchically structured document repositories.

4 Related Work

Systems such as Bead (Chalmers, 1993) and SPIRE (Thomas et al., 2001) map documents from a high-dimensional term space to a lower dimensional display space, whilst preserving the high-dimensional distances as far as possible. In contrast to InfoSky, both operate on flat document repositories and do not take advantage of hierarchical structure. Systems such as the Hyperbolic

Browser (Lamping, Rao, & Pirolli, 1995) and Information Pyramids (Andrews, Wolte, & Pichler, 1997) visualise large hierarchical structures while optimising the use of screen real estate, but make no explicit use of document content and subcollection similarities. CyberGeo Maps (Holmquist, Fagrell, & Busso, 1998) use a stars and galaxy metaphor similar to InfoSky, but the hierarchy is simply laid out in concentric rings around the root. WebMap's InternetMap (WebMap, 2002) visualises hierarchically categories of web sites recursively as multi-faceted shapes. However, unlike InfoSky, there is no correspondence between the local view at each level and the global view.

5 Concluding Remarks

This paper presented InfoSky, a first prototype system for the interactive visualisation and exploration of large, hierarchically structured, document repositories. Using its telescope and galaxy metaphor, the InfoSky system addresses several key requirements for such systems. As development proceeds, we believe that the InfoSky prototype will constitute a step towards practical, user-oriented, visual exploration of large, hierarchically structured document repositories. Readers are referred to detailed descriptions of both InfoSky and the user study in (Andrews et al., 2002).

References

Andrews, K., Kienreich, W., Sabol, V., Becker, J., Droschl, G., Kappe, F., Granitzer, M., Auer, P., & Tochtermann, K. (2002). The infosky visual explorer: Exploiting hierarchical structure and document similarities. *Information Visualization*, *1*(3/4), 166–181.

Andrews, K., Wolte, J., & Pichler, M. (1997). Information pyramids: A new approach to visualising large hierarchies. In *Ieee visualization'97, late breaking hot topics proc.* (pp. 49–52). Phoenix, Arizona.

Chalmers, M. (1993). Using a landscape metaphor to represent a corpus of documents. In *Spatial information theory, proc. cosit'93* (pp. 377–390). Boston, Massachusetts.

Holmquist, L. E., Fagrell, H., & Busso, R. (1998). Navigating cyberspace with cybergeo maps. In *Proc. of information systems research seminar in scandinavia (iris 21).* Saeby, Denmark.

Lamping, J., Rao, R., & Pirolli, P. (1995). A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In *Proc. chi'95* (pp. 401–408). Denver, Colorado.

Okabe, A., Boots, B., Sugihara, K., & Chiu, S. N. (2000). *Spatial tesselations: Concepts and applications of voronoi diagrams* (Second ed.). Wiley.

Thomas, J., Cowley, P., Kuchar, O., Nowell, L., Thomson, J., & Wong, P. C. (2001). Discovering knowledge through visual analysis. *Journal of Universal Computer Science*, *7*(6), 517–529.

WebMap. (2002). *WebMap.* (`http://www.webmap.com/`)