InfoSky: A System for Visual Exploration of Very Large, Hierarchically Structured Knowledge Spaces

Wolfgang Kienreich
Know-Center
Inffeldgasse 16c, 8020 Graz
wkien@know-center.at

Vedran Sabol
Know-Center
Inffeldgasse 16c, 8020 Graz
vsabol@know-center.at

Michael Granitzer
Know-Center
Inffeldgasse 16c, 8020 Graz
mgrani@know-center.at

Frank Kappe
Hyperwave R&D
Albrechtgasse 9, 8010 Graz
fkappe@hyperwave.com

Keith Andrews
IICM, TU-Graz
Inffeldgasse 16, 8020 Graz
kandrews@iicm.edu

Abstract

This publication presents InfoSky, a system enabling exploration of large, hierarchically structured knowledge spaces. InfoSky employs a two-dimensional graphical representation with variable magnification, much like a real-world telescope, to visualise individual documents as stars, hierarchical structures as constellations, and the whole knowledge repository as a galaxy. Force-directed placement is used to position topically similar documents in geometric adjacency, and modified Voronoi diagrams are employed to construct non-overlapping constellation boundaries, while statistical text processing extracts abstracts and keywords from documents and collections. InfoSky combines hierarchy and topical similarity in a visualisation using a striking, well-known metaphor, providing users with a tool appropriate for today's large, hierarchically structured document repositories.

1 Introduction

The problem of interactively visualising very large, hierarchically structured document collections, as well as visualising the results of retrieval operations executed on such collections has become a major research focus. Due to the ever-increasing number of entities stored within knowledge repositories like corporate intranets, hierarchical structures for organising documents into collections are more and more replacing flat repositories. Consequentially, users want both to browse large structured information spaces and be able to search them based on explicit criteria. However, many state-of-the-art retrieval and visualisation tools operate on flat, unstructured repositories.

In another recent development, graphical processing power on desktop systems has multiplied rapidly. Today, typical desktop computers are capable of visualising millions of entities in real time. Furthermore, interactive graphical representations of abstract information are becoming increasingly common and users are becoming familiar with such systems, resulting in reduced accessibility barriers and training requirements.

Based on this developments, the following requirements for a next generation document repository visualisation tool were formulated:

1. Scalability. Visualise very large (hundreds of thousands, if not millions of entities), hierarchically structured document repositories.

2. Hierarchy plus similarity. Combine both the hierarchical organisation of documents and inter-document similarity within a single, consistent visualisation.

3. Focus plus context. Integrate both a global and a local view of the information space into one seamless visualisation.

4. Query plus Exploration. Provide simple, intuitive facilities to browse and search the repository. Let the visualisation display a maximum number of properties and relationships to replace dedicated queries with exploration as much as possible.

The InfoSky system, shown in Figure 1, addresses the above requirements. InfoSky, the initial prototype of which was called KnowledgeScope, enables users to explore large, hierarchically structured document collections. Similar to a real-world telescope, InfoSky employs a two-dimensional graphical representation with variable magnification. 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.

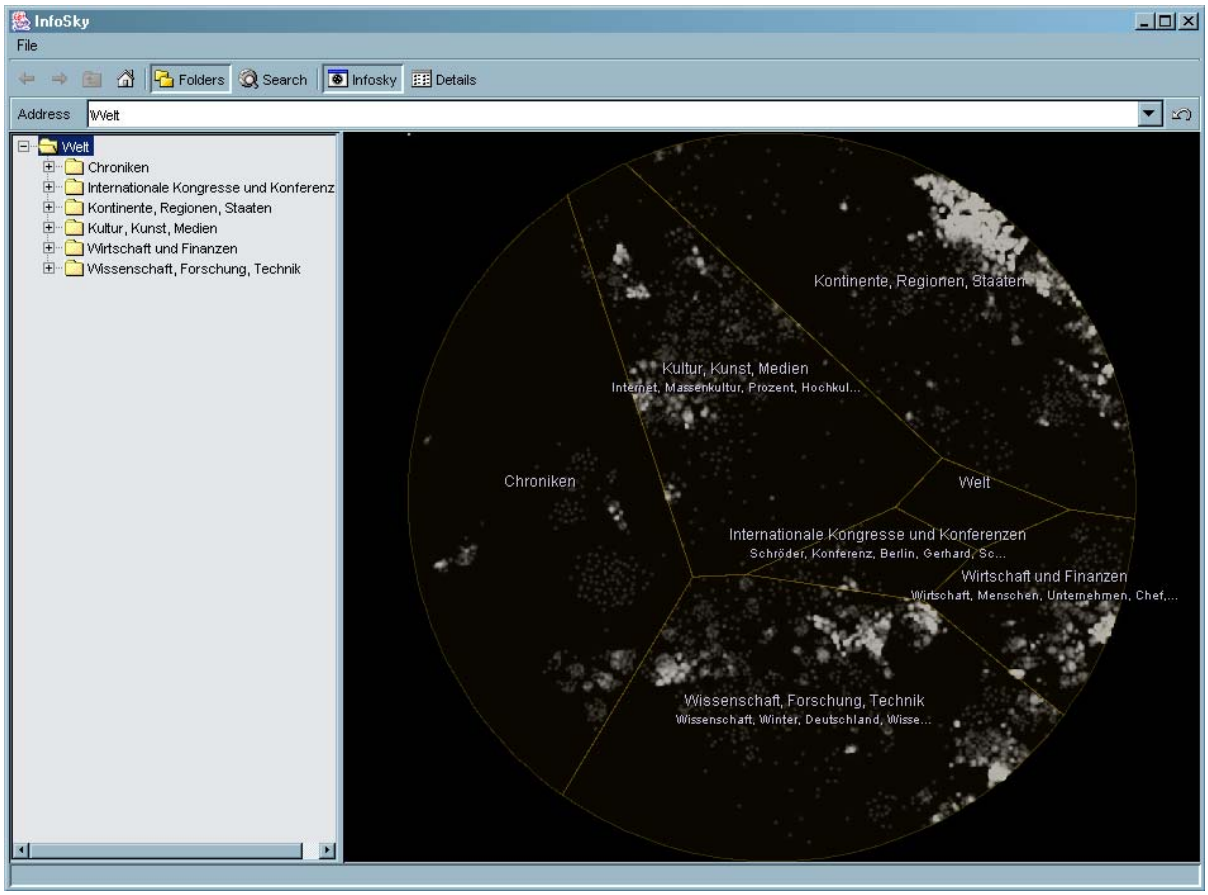

Figure 1: The InfoSky Visual Explorer

In this publication, we introduce InfoSky, and describe related work and similar systems. We provide an overview on the techniques and algorithms used in implementing the system, and give results of usability studies done so far. Finally, we provide an outlook on future activities.

2 InfoSky

InfoSky uses a zooming galaxy of stars, organised hierarchically into clusters, as a metaphor, with stars representing documents and constellations representing collections. The system assumes that documents are already organised in a hierarchy of collections and sub-collections, called the collection hierarchy. This is the case in most knowledge management systems. Both documents and collections can be members of more than one parent collection, the only restriction being that no cycles are allowed. The collection hierarchy might, for example, be a classification or categorization scheme, which could be manually maintained by editorial staff, or automatically created by another system. Documents are assumed to have significant textual content, with typical formats being plain text, PDF or HTML. Access to both documents and collections may be restricted according to assigned user rights: InfoSky takes such restrictions into account when displaying the visualisation. Depending on their access rights, certain users may not be able to view particular documents or collections. Meta-information present in the repository, such as author and modification date, can be processed and visualised by the InfoSky system, but the core visualisation is generated from actual document content.

2.1 Metaphor and Interface

InfoSky combines both a traditional tree browser and a new telescope view of a galaxy. In the galaxy, documents are visualised as stars, with similar documents forming 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. Magnification

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

Several facilities support users in operating the visualisation. Simple interactions cause the system to automatically shift focus to an object of interest and magnify it to optimal viewing size. When changing the magnification or position manually, constellation boundaries are automatically displayed and hidden to avoid display cluttering. Finally, history and bookmark functions allow easy recall of previously visited "galactic coordinates".

The right hand side of Figure 1 shows a galaxy view in InfoSky. This galaxy is derived from a collection of approximately 109,000 German language news articles from the German daily newspaper, the Süddeutsche Zeitung. The articles have been classified thematically by the newspaper's editorial staff into around 6,900 collections and sub-collections up to 15 levels deep. Constellation boundaries and labels are shown for the topmost level of the hierarchy.

The galaxy itself is complete in the sense that it displays all the stars it contains, down to the bottom-most level of the hierarchy. At this level of magnification, individual stars are not discernible. The clusters forming the galaxy consist of thousands of stars which, in accordance with the telescope metaphor, can only be resolved individually at a higher magnification.

2.3 Navigation

Interactive exploration of the galaxy is supported by a combination of browsing and searching capabilities. Selection of a region of interest (a collection or document) causes that region to be auto-centred: the viewport and magnification are adjusted so that the region of interest is displayed in full. In addition, the user can freely change the current view by changing the magnification (zooming) and sliding the viewport around at the current magnification (panning). While zooming and panning, collections are auto-selected based on magnification and position: the maximum level of the hierarchy fitting completely inside the viewport is determined and the collection at that level nearest to the centre of the viewport is selected. To ensure the widest possible audience, only a keyboard and mouse are used for navigation. Furthermore, the standard tree view is fully synchronised with the visualisation, allowing new users to slowly switch from their common navigation mode to the new features of InfoSky.

Figure 2 displays part of the galaxy whose global view is shown in Figure 1. Still, no resolution into individual stars has been reached, however, a subset of the constellations of Figure 1 is now visible. Figure 3 displays the galaxy after zooming further into the collection displayed in Figure 2, finally providing resolution of clusters into individual stars. The collection has been centred and magnification adjusted to make the collection fill the viewport.

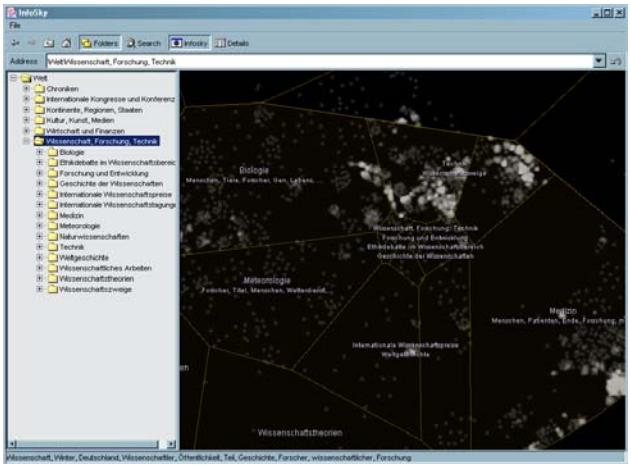

Figure 2: Zooming into the Galaxy

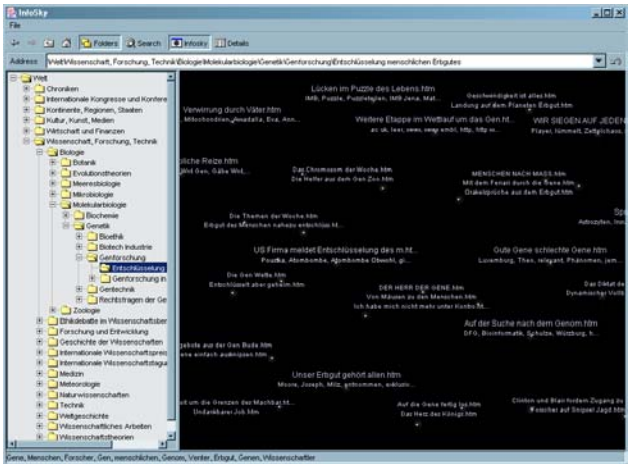

Figure 3: Resolving individual stars

2.4 Searching

Users can search for documents and collections contained in the corpus by issuing a query. Matching documents and collections are highlighted and can be examined in further detail. Figure 4 shows matching documents corresponding to the query "virus". For experienced analysts, a further query mode allows the results of individual queries to be assigned to an individual colour channel and overlays created to express the combined results of several queries.

In addition to InfoSky's search result visualisation, a standard relevance ranked list view of is also available. This view displays ranking, keywords and a brief abstract for each found document and is fully synchronised with the galaxy view.

3 Related Work

Published work on the visualisation of large document repositories concentrates on approaches utilising inter-document similarity measures within flat repositories on one hand, and on visual exploration of hierarchically organised structures on the other. Only recently have some first steps been taken towards integrating these two approaches.

3.1 Approaches Based on Similarity

A number of systems employ methods for mapping documents from a high-dimensional term space to a lower dimensional display space, whilst preserving the high-dimensional distances as far as possible. The Bead system [Chalmers, 1993, Chalmers, 1996] is a typical example of a thematic landscape. The information space is arranged based on inter-document similarity forming a 2.1D landscape. Users can navigate freely around the information landscape, with search result being displayed directly within the landscape, too. In contrast to InfoSky, Bead operates on flat document repositories and does not take advantage of hierarchical structure.

Galaxy Of News [Rennison 1994] constructs and then visualises an associative relation network between related news articles. At first, a hierarchy of topical keywords from general to more specific is presented, which then lead into article headlines, and eventually to full news articles. Unlike InfoSky, the space is non-linear and changes as the user navigates, making it hard to maintain a sense of orientation.

SPIRE [Thomas *et al.*, 2001] operates on flat, unstructured document collections. SPIRE's Galaxies visualise documents as stars in a galaxy, where documents which are close in high dimensional space are also close in the two-dimensional galaxy view. This is similar to the approach taken by InfoSky to lay out documents at any particular level of the collection hierarchy. SPIRE does not exploit any inherent hierarchical structure.

Earlier work at the IICM on VisIslands [Andrews *et al.*, 2001; Sabol, 2001] used standard clustering techniques to cluster document sets returned in response to a search query on the fly.

WEBSOM [WEBSOM, 2000] and other systems employ self-organising maps to thematically organise and visualise very large document collections. However, the underlying neural networks have to undergo extensive training in order to achieve good results.

3.2 Approaches Based on Hierarchical Structure

Systems focusing on the visualisation and navigation of large hierarchical structures tend to optimise the use of screen (pixel) real estate by either (or both) geometric transformations or zooming and panning interactions.

The Hyperbolic Browser [Lamping *et al.*, 1995] is a two-dimensional tree browser, which utilises hyperbolic geometry to always display the entire hierarchy on the display. The tree is laid out using hyperbolic axes (which are infinite) and then mapped to a two-dimensional unit disc for display. Areas in the centre of the disc are in focus and clearly visible, areas toward the edge of the disc become infinitely small and can no longer be seen.

The H3 browser [Munzner, 1997] makes even better use of screen space by using 3D, at the cost of some occlusion. However, neither of these systems make explicit use of document content and sub-collection similarities.

Cone Trees [Robertson *et al.*, 1991] lay out hierarchies in three dimensions. Each node in the hierarchy is the apex of a cone. The root of the hierarchy is placed near the top of the three-dimensional display space and its children are evenly spaced along its base. The next layer of nodes is drawn below the first layer, recursively until the whole hierarchy is drawn. Nodes are usually given a label or name. Cone trees suffer from problems of occlusion as hierarchies become broad and branches become hidden behind their siblings, interactivity has to be employed to rotate hidden branches.

CyberGeo Maps [Skog *et al.*, 2000] use a stars and galaxy metaphor to lay out pages of a web site. A manually edited hierarchical categorisation is composed, roughly corresponding to the directory structure of the web site. The root of the hierarchy corresponds to the sun at the centre of the solar system. Dots (stars) representing web pages are placed at orbits around the centre, depending on how far away they are from the home page. The metaphor is similar to that used in InfoSky and the visual display is similar, but the underlying layout is very different. In CyberGeo Maps levels of the hierarchy form concentric rings around the root, in InfoSky Voronoi diagrams are used recursively. In CyberGeo Maps the proximity of stars is not related to their similarity.

3.3 Integrated Approaches

Information Pyramids [Andrews *et al.*, 1997] use a three-dimensional landscape to visualise a hierarchy. Full usage of the third dimension is made by visualising both the content and structural information in three dimensions. The general impression is that of pyramids growing upwards as the hierarchy grows deeper. Whereas Information Pyramids uses recursive placement of rectangles at each level of the hierarchy, InfoSky uses recursive partition of polygons with Voronoi diagrams.

WebMap's InternetMap [Iron *et al.*, 2001] visualises hierarchically categorised web sites. Each site is represented by a pixel, sites belonging to multiple categories are represented by separate pixels in each category. Each category is visualised as a multi-faceted shape, enclosing the sites within that category. Within a category, sites with similar content are geometrically close. However, there is no correspondence between the local view at each level and the global view.

4 Techniques

InfoSky consists of a server and a number clients. 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. Together, these components are able to generate a galaxy representation from millions of documents within a few hours, and to visualise the galaxy in real time on a standard desktop computer.

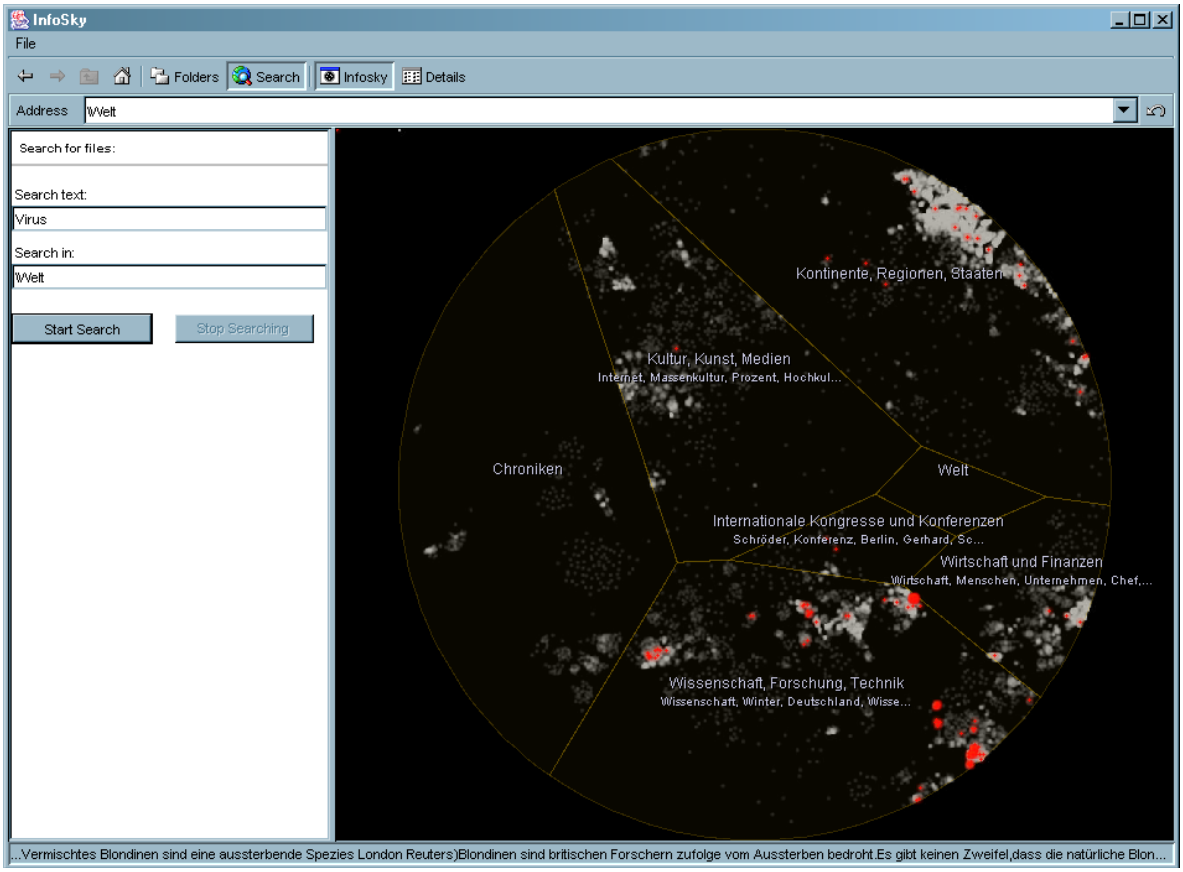

Figure 4: Search results for "Virus" being displayed (red stars)

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 similarity placement algorithm. The similarities to their parent's sibling collection centroids are used as static influence factors to ensure that similar neighbouring sub-collections across collection boundaries tend towards each other. The centroid of a synthetic sub-collection 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 sub-collection centroid, whose size is related to the total number of documents and collections contained in that sub-collection (at all lower levels). This polygonal partition of the parent collection's area is done with a modified Voronoi diagram [Okabe *et al.*, 2000].

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

Three algorithms are particularly prominent:

1. Similarity placement: Similarity placement is used to position both sub-collection centroids within their parent collection and to position documents within the synthetic Stars collection. Similarity placement is realised using an optimised force-directed placement algorithm.

2. Geometric transformation: Document and collection centroid positions are defined in a normalised space after similarity placement has been done. However, final placement has to be done within a bounding polygon. Consequentially, coordinates are transformed from the normalised coordinate system into a multi-axis coordinate system formed by the bounds.

3. Area partition: The centroids of sub-collections are used to partition the polygon representing the parent collection into polygonal sub-areas. The size of each sub-area is related to the total number of documents contained within the corresponding sub-collection. Area partition is accomplished using modified, weighted Voronoi diagrams.

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 sub-collection centroids plus the number of documents at that level).

5 Evaluation

Some initial formative thinking aloud testing of the entire InfoSky prototype was done with colleagues at the Know-Center. It was then decided to run a first formal experiment to establish a baseline comparison between the InfoSky telescope browser and the InfoSky tree browser. Due to users' much greater familiarity with Windows Explorer style tree browsers and the early development nature of the telescope browser prototype, it was expected that users would probably be more efficient initially using the tree browser alone than the prototype telescope browser alone, but running a study would provide a feeling for how much of a difference there was. Note that this initial study did not test the power of a combination of tree browser, telescope browser, and search functionality.

The dataset from the Süddeutsche Zeitung was taken and two sets of tasks were formulated. 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. A pilot test with one user was run and some slight modifications were made. Then, the real test was executed.

On average, the tree browser performed better than the prototype telescope browser for each of the tasks tested. 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. They were not familiar with the telescope browser and two minutes of training could not make up the deficit. 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 and has numerous bugs and issues affecting its usability. Some of the problems in the current implementation of the telescope browser which emerged during the study were:

1. The Voronoi polygons in the centre of each collection were far too small for many test users and a minimum size should probably be set.

2. When near the bottom of the hierarchy, where collections contained many documents, users were confused by the "jumping around" of document titles. The prototype displayed the titles of those documents which were "near" to the cursor.

3. When more than a handful of document titles were displayed, the telescope display became cluttered. Problems 2 and 3 might be alleviated by displaying a scrolling, linear list of documents once users reach a level of the hierarchy where they want to inspect document titles.

4. The synthetic collection "Stars" containing documents at a particular level of the collection hierarchy was confusing to users.

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 sub-collections, 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.

User feedback has already been taken into account in the latest version of the InfoSky system. For example, Figure 3 displays that the overlapping and jumping of labels as described in problem 3 has been eliminated by using an advanced label merging and layouting algorithm which guarantees that no overlaps occur: Whenever two labels tend to overlap, they are merged into a single, new label.

6 Outlook

In this publication, we have presented InfoSky, a system for visual exploration of large, hierarchically structured knowledge spaces. InfoSky is constantly being revised and expanded in close cooperation with Hyperwave R&D, who has filed patent for the system [Kappe *et al.*, 2001]. Upcoming challenges include full integration into a knowledge-management system, the Hyperwave E-knowledge Suite, incorporation of multimedia documents, enhancements to the user interface and further usability testing.

References

[Abelson *et al.*, 1985] Harold Abelson, Gerald Jay Sussman, and Julie Sussman. *Structure and Interpretation of Computer Programs*. MIT Press, Cambridge, Massachusetts, 1985.

[Abelson *et al.*, 1985] Harold Abelson, Gerald Jay Sussman, and Julie Sussman. *Structure and Interpretation*

of Computer Programs. MIT Press, Cambridge, Massachusetts, 1985.

[Andrews *et al.*, 2001] Keith Andrews, Christian G¨utl, Josef Moser, Vedran Sabol, and Wilfried Lackner. *Search result visualisation with xfind*. Proc. UIDIS 2001, pages 50–58, Zurich, Switzerland, May 2001. IEEE Computer Society Press.

[Andrews *et al.*, 1997] Keith Andrews, Josef Wolte and Michael Pichler. *Information pyramids: A new approach to visualising large hierarchies*. IEEE Visualization'97, Late Breaking Hot Topics Proc., pages 49–52, Phoenix, Arizona, October 1997

[Benderson *et al*., 1994] Ben Bederson and Jim Hollan. *Pad++: A zooming graphical interface for exploring alternative interface physics*. Proc. UIST'94, pages 17–26, Marina del Rey, CA, November 1994. ACM.

[Chalmers 1993] Matthew Chalmers. *Using a landscape metaphor to represent a corpus of documents*. Spatial Information Theory, Proc. COSIT'93, pages 377–390, Boston, Massachusetts, September 1993. Springer LNCS 716.

[Chalmers 1996] Matthew Chalmers. *A linear iteration time layout algorithm for visualising high-dimensional data.* Proc. Visualization'96, pages 127–132, San Francisco, California, October 1996. IEEE Computer Society.

[Holmquist *et al.*, 1996] Lars Erik Holmquist, Henrik Fagrell, and Roberto Busso. *Navigating cyberspace with cybergeo maps.*. Proc. of Information Systems Research Seminar in Scandinavia (IRIS 21), Saeby, Denmark, August 1998

[Iron *et al.*, 2001] Michael Iron, Roi Neustedt, and Ohad Ranen. *Method of graphically presenting network information*. US Patent Application 20010035885A1, WebMap, November 2001. Filed March 2001.

[Kappe *et al.*, 2001] Frank Kappe, Vedran Sabol, Wolfgang Kienreich. *InfoSky*. US Patent Application 60/376, Hyperwave. Filed May 2002. International Patent Application PCT/EP03/03445, Hyperwave. Filed July 2002.

[Lamping *et al.*, 1995] John Lamping, Ramana Rao, and Peter Pirolli. *A focus+context technique based on hyperbolic geometry for visualizing large hierarchies.* Proc. CHI'95, pages 401–408, Denver, Colorado, May 1995. ACM.

[Munzner 1995] Tamara Munzner. *H3: Laying out large directed graphs in 3d hyperbolic space.* Proc. IEEE InfoVis'97, pages 2–10, Phoenix, Arizona, October 1997

[Okabe *et al.*, 2000] Atsuyuki Okabe, Barry Boots, Kokichi Sugihara, and Sung Nok Chiu. *Spatial Tesselations: Concepts and Applications of Voronoi Diagrams.* Wiley, second edition, 2000. ISBN 0471986356.

[Rennison, 1994] Earl Rennison. *Galaxy of news: An approach to visualizing and understanding expansive news landscapes.*. Proc. UIST'94, pages 3–12, Marina del Rey, California, November 1994. ACM

[Robertson *et al.*, 1991] George G. Robertson, Jock D. Mackinlay, and Stuart K. Card. *Cone trees: Animated 3D visualizations of hierarchical information.* In Proc. CHI'91, pages 189–194, New Orleans, Louisiana, May 1991. ACM.

[Sabol 2001] Vedran Sabol. *Visualisation islands: Interactive visualisation and clustering of search result sets.*. Master's thesis, Graz University of Technology, Austria, October 2001.

[Skog *et al.*, 2000] Tobias Skog and Lars Erik Holmquist. *Continuous visualization of web site activity in a public place.* Student Poster, CHI 2000 Extended Abstracts, The Hague, The Netherlands, April 2000.

[Strasnick *et al.*, 1996] Steven L. Strasnick and Joel D. Tesler *Method and apparatus for displaying data within a threedimensional information landscapeplace.* US Patent 5528735 , Silicon Graphics, Inc., June 1996. Filed 23rd March 1993, issued 18th June 1996.

[Thomas *et al.*, 2001] Jim Thomas, Paula Cowley, Olga Kuchar, Lucy Nowell, Judi Thomson, and Pak Chung Wong *Discovering knowledge through visual analysis.* Journal of Universal Computer Science, 7(6):517–529, June 2001

[WebSom., 2000] *Websom - self-organizing maps for internet exploration*. Helsinki University of Technology 2000