Evaluating a System for Interactive Exploration of Large, Hierarchically Structured Document Repositories

Michael Granitzer[+] Know-Center Graz
Wolfgang Kienreich[+] Know-Center Graz
Vedran Sabol[+] Know-Center Graz
Keith Andrews[*] Technical University Graz
Werner Klieber[+] Know-Center Graz

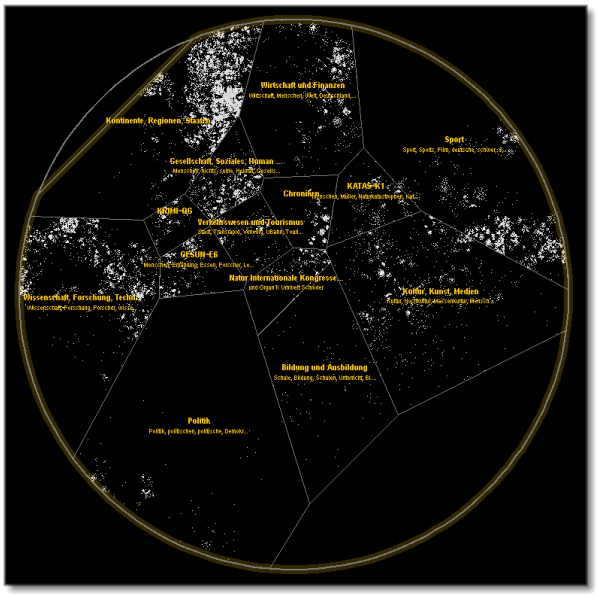

Figure 1: The InfoSky Visualisation

ABSTRACT

The InfoSky visual explorer is a system enabling users to interactively explore large, hierarchically structured document collections. Similar to a real-world telescope, InfoSky employs a planar graphical representation with variable magnification. Documents of similar content are placed close to each other and displayed as stars, while collections of documents at a particular level in the hierarchy are visualised as bounding polygons.

Usability testing of an early prototype implementation of InfoSky revealed several design issues which prevented users from fully exploiting the power of the visual metaphor. Evaluation results have been incorporated into an advanced prototype, and another usability test has been conducted. A comparison of test results demonstrates enhanced system performance and points out promising directions for further work.

CR Categories: H.3.3 [Information Search and Retrieval]: Search Process; I.3.6 [Methodology and Techniques]: Interaction Techniques

Keywords: information visualisation, navigation, document retrieval, hierarchical repositories, knowledge management, information management, force-directed placement, Voronoi.

[+] email: {mgrani|wkien|vsabol|wklieber}@know-center.at
[*] email: kandrews@iicm.edu

1 INTRODUCTION

In reaction to the steadily growing amount of information in corporate intranets as well as on the word-wide web, structuring is applied to an increasing number of document repositories. Unfortunately, the deep classification hierarchies used to organize document collections containing millions of items cannot be easily navigated and searched using existing visual retrieval tools, which are often tailored towards flat repositories containing several thousands of documents at most. Important concepts of information visualisation, like seamless transition between overview and detail view, do not easily scale to the amount of data future repositories are likely to contain, and many metaphors working well for today's flat repositories cannot easily be applied to hierarchies.

The InfoSky visual explorer has been designed and implemented with these challenges in mind. A patented method exploits hierarchical structure for performance optimisation, generating a similarity-based 2D-layout for millions of documents in thousands of collections. The night sky is used as a visualisation metaphor, and user interaction is designed around the idea of providing a virtual telescope. Employing these concepts, InfoSky addresses the following main challenges:

- Hierarchy plus similarity: Represent both the hierarchical organisation of documents and inter-document similarity within a single, consistent visualisation.
- Focus plus context: Integrate both a global and a local view of the information space into one seamless visualisation.
- Stability: Use a stable metaphor which promotes visual recall and recognition of features. The visualisation should remain

largely unchanged at a global level even if changes occur to the underlying document repository on a local level.

- Unified frame of reference: Support a single, consistent view of the document space for all users, regardless of the access rights of each individual user, thus providing a common frame of reference for all parties.
- Exploration: Provide simple, intuitive facilities to browse and search the repository. The visualisation tool should allow the visualisation to display a maximum number of document properties and relationships without any need for user interaction. It should thus offer a means of locating documents without specifying a query, by simply browsing the information space and displaying information within its context.
- Scalability: Visualise very large (hundreds of thousands, if not millions of entities), hierarchically structured document repositories.

This publication presents the InfoSky visual explorer and discusses recent evaluation result obtained from usability studies carried out using an advanced system prototype. Section 2 presents the philosophy and interface of the InfoSky visual explorer. Section 3 reviews the initial evaluation of a first system prototype done in 2002. Section 4 discusses the results of the recent evaluation of an enhanced prototype which has been built based on the 2002 results. Section 5 discusses related work and Section 6 describes possible next steps in the development of InfoSky.

2 INFOSKY

InfoSky employs the metaphor of an interactively zooming galaxy of stars, organised hierarchically into recognizable thematical clusters. The underlying data source is assumed to be a hierarchically structured document repository, where document collections and sub-collections form a directed acyclic graph in which both documents and collections can be assigned to more than one parent collection. The collection hierarchy might, for example, be a classification scheme or taxonomy, manually maintained by editorial staff or generated (semi-)automatically.

Documents are assumed to have significant textual content, which can be extracted and processed to provide measures for inter-document similarity. Typical document formats include text, PDF, HTML, or Word. Access to both documents and collections can be restricted according to assigned user rights, resulting in inaccessible documents and collections being hidden from users. Meta-information present in the repository, such as author and modification date, can also be incorporated and visualised by the system, but the actual visualisation is generated mainly from the document content.

2.1 The Telescope Metaphor

InfoSky integrates both a traditional tree browser and the 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 very deep in the hierarchy, down to the level of clusters and stars, or reduced to display the galaxy as a whole. Several facilities support users in operating this virtual telescope. 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”.

2.2 Navigating the Galaxy

Interactive exploration (navigation) of the galaxy is achieved through 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 address the widest possible audience, only a keyboard and mouse are used for navigation. In the current prototype, the following navigational facilities are provided (note that these can easily be changed and extended):

- Selecting a collection: Left-clicking a collection label selects the collection and auto-centres it.
- Selecting a document: Left-clicking an individual star selects the corresponding document and auto-centres it.
- Selecting the parent collection: A toolbar button allows to place the focus on the parent collection. The viewport is zoomed out to display the collection.
- Continuous hierarchical zoom: After selecting a collection, zooming in on the visualisation continuously selects deeper hierarchical levels based on magnification and position.
- Panning: Dragging with the left mouse button pans the viewport. Collections are auto selected based on magnification and position.
- Zooming: Using the mouse-wheel, the magnification factor of the display can be adjusted.

The many features supporting interaction are very important for intuitive navigation. In particular, continuous hierarchical zoom represents a significant advance over conventional step-by-step browsing of a hierarchy. Similar to related work on zooming interfaces by Bederson and Hollan [2][3], continuous hierarchical zoom allows users to bypass upper levels of the hierarchy and quickly move to a known position within the galaxy. Without continuous zoom, users must explicitly select the correct child collection at each hierarchy level, until the desired collection is reached, resulting in a greatly increased number of interactions comparable to the conventional tree browser. The usability of the continuous hierarchical zooming facilities in InfoSky has been enhanced since the initial version by automating the process: Hierarchical zoom occurs automatically whenever viewport size, location or zoom level changes.

2.3 Searching for Documents

InfoSky features sophisticated search functionality, including the ability to execute a number of independent queries. Results of each query are displayed as color-coded stars representing found documents. By using a different colour for every displayed query results can be combined making the degree of overlapping immediately visible. One benefit of visualising search results in InfoSky is that the context of a given result item is immediately clear, and similar results which have not been covered by the search are located close to the result item. However, the usability experiments discussed in this publication did not test InfoSky's search facilities, this will be done in a separate study.

2.4 Implementation

Only a brief overview of the implementation details of InfoSky is given in this publication. All algorithms used have been described in detail in the pending patent [1]. For a comprehensive scientific presentation, please refer to [4].

InfoSky is implemented as a client-server system. 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. Java was chosen as the development platform for both client and server, because of its platform-independence and geometric libraries. 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. The galactic geometry is generated from the underlying repository recursively from top to bottom in several steps.

First, at each level or the hierarchy, the sub-collection centroids are positioned in a normalised 2D plane according to their similarities using a similarity placement algorithm. The similarities to their parent's sibling collections are used as static influence factors to ensure that similar neighbouring sub-collections across collection boundaries tend towards each other (they are not allowed to actually cross the boundary). The centroid of a synthetic sub-collection called "Stars", which holds the documents at that level of the hierarchy, is also positioned together with the sub-collections. Similarity placement is realised using an optimised force-directed placement algorithm [5]. The layout in normalised 2D space is transformed to the polygonal area of the parent collection using a simple geometric transformation.

Then, 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 [6].

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 interdocument 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. Force-directed placement (FDP) is an iterative method for mapping a set of high-dimensional vectors to a low-dimensional space, whilst preserving their high-dimensional relations as far as possible. The algorithm calculates force vectors from the similarities between documents and collection centroids. These forces, and additional, custom-defined vectors, influence the position of the objects at each iteration in the placement algorithm.
2. Geometric transformation: The geometric transformation employed inscribes all points into the bounding polygon of the collection using a simple geometric transform.
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.

The use of Voronoi diagrams to represent the hierarchical structure of the underlying repository introduced several problems relating to the varying size of collections. In a standard Voronoi diagram, available space is evenly distributed between all participating points. In order to represent the number of documents and sub-collections contained in a given collection, additively weighted power Voronoi diagrams have been used, and some modifications have been made to the force-directed placement algorithm to assign more space to heavy-weight collections and to pull light-weight collection centroids towards the center of the parenting collection. As a result, the Voronoi partitions reflect collection sizes well in most cases (compare figure 2).

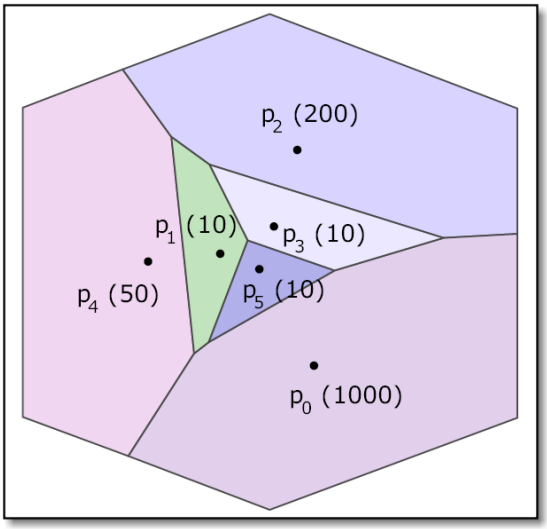

Figure 2: Voronoi layout of collection centroids

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).

3 Initial Study

The first prototype of InfoSky was evaluated in a formal experiment in 2002, to establish a baseline comparison between the InfoSky telescope browser and the InfoSky tree browser. The browser used is shown in Figure 5. Users were only allowed to use one or the other part of it in isolation. The browser was used in full screen mode, and the search box was removed.

The test dataset (consisting of 110.000 newspaper articles from the German Sueddeutsche Zeitung) was taken and 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. The test environment was set up as shown in Figure 3.

Figure 3: Test Setup in initial study

Eight employees of Hyperwave R&D were recruited for the study and divided randomly into four groups of two. Four users began with the telescope browser (condition TS), then used the tree view (condition TV). The other four users began with TV then used TS. Within these conditions two users started with task set A, the other two with task set B. Before using the telescope browser, users were given two minutes of brief training on the browser's features. At the end of each test, an interview was conducted with the test user to gain additional feedback. The entire session was videotaped.

	TP1	TP2	TP3	TP4	TP5	TP6	TP7	TP8	Av	Diff	Diff %
TV1	7,0	12,0	2,0	5,0	11,0	63,0	15,0	7,0	15,3		
TV2	8,0	6,0	4,0	9,0	7,0	6,0	8,0	14,0	7,8		
TV3	69,0	40,0	42,0	84,0	13,0	34,0	23,0	137,0	55,3		
TV4	8,0	8,0	7,0	13,0	32,0	41,0	6,0	14,0	16,1		
TV5	32,0	37,0	75,0	85,0	55,0	43,0	32,0	18,0	47,1		
TS1	21,0	24,0	10,0	40,0	9,0	7,0	22,0	33,0	20,8	5,5	27%
TS2	15,0	9,0	11,0	22,0	3,0	15,0	11,0	7,0	11,6	3,9	33%
TS3	189,0	114,0	35,0	121,0	32,0	71,0	74,0	22,0	82,3	27,0	33%
TS4	32,0	94,0	35,0	16,0	84,0	48,0	39,0	21,0	46,1	30,0	65%
TS5	148,0	72,0	52,0	143,0	50,0	57,0	36,0	194,0	94,0	46,9	50%

Figure 4: Results of initial study

The results of the initial study are summarised in Figure 4. Timings were determined by analysing the videotape of each session and noting the time in seconds from the time the facilitator read the last word of the task to the time the task was completed. The overall difference between tree browser and telescope browser was significant at $p < 0.05$ (paired samples t-test, 39 degrees of freedom, $t = 3.038$), with the tree browser performing better than the prototype telescope view on average. Leaving aside the lack of familiarity of users with the telescope view, the main reason for the difference seemed lie in several implementation flaws of the telescope view:

- The Voronoi polygons in the centre of each collection were far too small for many test users
- 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.
- When more than a handful of document titles were displayed, the telescope display became cluttered.
- 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. Users further indicated that a combination of both browsers and search functionality could be very powerful.

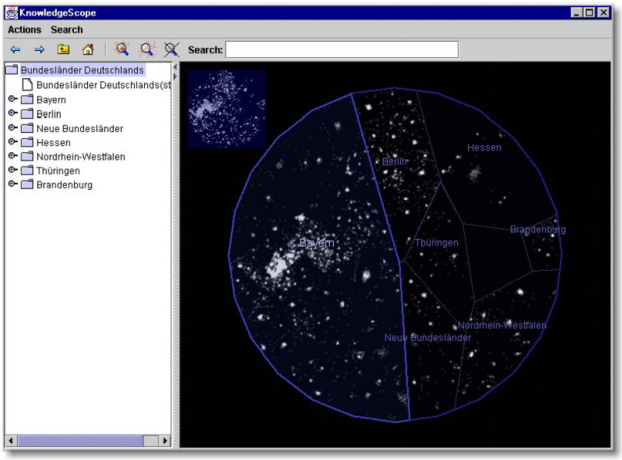

Figure 5: Prototype as tested in initial study

The findings of this baseline evaluation have been taken into account, and an extensive redesign phase, followed by another user test has been laid out.

4 Current Study

After one year of further developing InfoSky and adjusting it to the results of the initial usability study, a second evaluation and test was planned and executed in spring 2004. The user interface has been revised to match user feedback. For example, a list of documents in the selected collection has been added, layouting of labels in the galaxy view has been revised to minimize cluttering, and several interactions as described in 2.2 have been optimized based on user feedback received during the initial study. The resulting prototype used in the recent experiments is shown in Figure 6. Note that the user interface is divided into three distinct areas, the tree view (top-left), the galaxy view (top-right) and the list view (bottom). Combinations of tree and galaxy view have

been tested in the experiment, with the list view remaining in place.

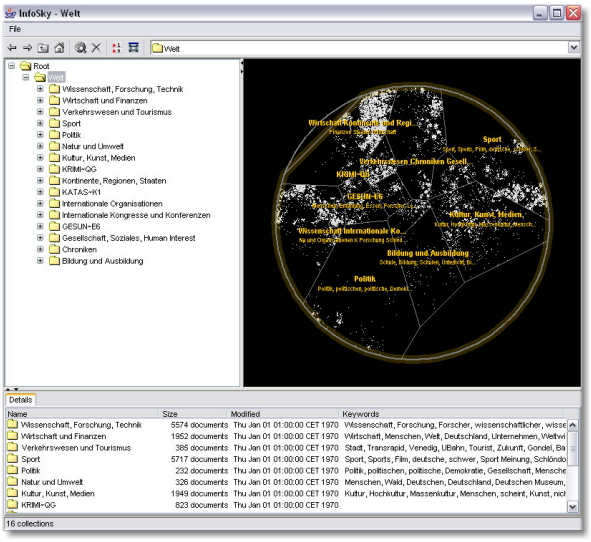

Figure 6: Revised prototype used in current study

The browser was evaluated in full screen mode, with all search functionality disabled in the toolbar. In contrast to the initial study, however, a test of the combination of tree and galaxy view was incorporated.

4.1 Test Setup

The test dataset (consisting of 80.000 newspaper articles from the German Sueddeutsche Zeitung) was taken and three sets of tasks were formulated (six triples of equivalent tasks). The tasks were designed to be equivalent among the three 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. The test environment setup followed the one used in the initial evaluation (compare section 3).

Nine employees of the Know-Center were recruited for the study and divided randomly into three groups of three. Users of the first group began with the telescope browser (condition GV), then used the tree view (condition TV), and finally the mixed view (condition MV) displaying both the tree and the telescope view . The other two groups used alternating ordering of test conditions. Before using the telescope browser (either stand-alone or in conjunction with the tree view), users were given a brief training on the browser's features. At the end of each test, an interview was conducted with the test user to gain additional feedback. The entire session was videotaped.

Questions included locating a document or collection within the hierarchy, counting the number of documents contained within a collection, comparing the number of items contained within two separate collections and counting the number of similar documents existing in the same collection for a given document. Timings were determined by analysing the videotape of each session and noting the time in seconds from the time the facilitator read the last word of the task to the time the task was completed. Note that for several tasks, time-out occurred when a test user decided that he or she could not solve a given task and advanced to the next task. All time-outs have been left out of the statistic evaluation of results. The results of the initial study are summarised in Figure 6. The differences between the three test conditions were significant, the statistical analysis is given in Figure 7.

	TP1	TP2	TP3	TP4	TP5	TP6	TP7	TP8	TP9	Av	Diff	Diff %
TV1	14,0	11,0	9,0	9,0	11,0	5,0	6,0	6,0	9,0	8,9		
TV2	15,0	106,0	35,0	98,0	T	16,0	T	T	T	54,0		
TV3	163,0	212,0	61,0	79,0	246,0	51,0	78,0	123,0	134,0	127,4		
TV4	45,0	29,0	40,0	28,0	60,0	33,0	8,0	8,0	T	31,4		
TV5	135,0	63,0	T	49,0	33,0	62,0	94,0	40,0	T	68,0		
TV6	84,0	91,0	46,0	178,0	78,0	76,0	54,0	104,0	108,0	91,0		
GV1	34,0	56,0	25,0	39,0	49,0	21,0	31,0	15,0	117,0	43,0	34,1	79%
GV2	37,0	T	62,0	100,0	T	71,0	41,0	T	77,0	64,7	10,7	16%
GV3	68,0	273,0	414,0	107,0	400,0	400,0	141,0	420,0	T	277,9	150,4	54%
GV4	T	98,0	76,0	263,0	56,0	96,0	38,0	141,0	123,0	111,4	80,0	72%
GV5	70,0	105,0	169,0	T	130,0	T	63,0	133,0	T	111,7	43,7	39%
GV6	106,0	172,0	153,0	132,0	232,0	144,0	148,0	331,0	217,0	181,7	90,7	50%
MV1	16,0	17,0	12,0	14,0	20,0	7,0	16,0	8,0	27,0	15,2	6,3	42%
MV2	223,0	35,0	T	165,0	17,0	175,0	58,0	95,0	195,0	120,4	66,4	55%
MV3	119,0	122,0	78,0	321,0	58,0	161,0	128,0	169,0	T	144,5	17,1	12%
MV4	26,0	40,0	14,0	179,0	68,0	34,0	42,0	95,0	215,0	79,2	47,8	60%
MV5	54,0	267,0	62,0	78,0	106,0	99,0	117,0	256,0	178,0	135,2	67,2	50%
MV6	119,0	92,0	121,0	160,0	T	145,0	91,0	T	212,0	134,3	43,3	32%

Figure 7: Results of current study

	p	T		p	T		p	T
TV1-GV1	0,001	3,416	TV1-MV1	0,007	3,575	GV1-MV1	0,001	3,299
TV2-GV2	0,739	2,425	TV2-MV2	0,235	1,480	GV2-MV2	0,739	3,483
TV3-GV3	0,290	2,578	TV3-MV3	0,709	0,389	GV3-MV3	0,290	1,889
TV4-GV4	0,068	2,644	TV4-MV4	0,188	1,458	GV4-MV4	0,068	1,241
TV5-GV5	0,903	0,833	TV5-MV5	0,123	1,791	GV5-MV5	0,903	0,781
TV6-GV6	0,248	3,541	TV6-MV6	0,037	2,674	GV6-MV6	0,248	1,277

Figure 8: Statistical analysis of current study

4.2 Interpretation of results

It is important for the following analysis to note that the tasks given to test users in the recent usability study were much more sophisticated than in the initial evaluation. Tasks demanding the location of items referred to items deeper in the hierarchy, and several types of tasks (i.e. comparing the number of items in two collections or finding similar items) were not part of the initial study at all. With this in mind, the most important results found can be summarized:

- A combination of the tree browser and the galaxy browser (mixed mode) yields significantly better results than the use of the galaxy browser alone. This result does not only show up clearly in the statistical analysis, but is also underlined by comments given by users in the follow-up interviews. Users consistently emphasised the value of the telescope view as an overview tool which prevented them from getting lost when navigating deep within the hierarchy.
- While not directly comparable, the difference between mixed mode and the tree view in stand-alone mode approach the difference between the telescope view and the tree view found in the initial evaluation. In the light of the much more complex tasks given to users in the recent evaluation, this can be interpreted as a consequence of adaptations made based on the results of the initial evaluation.
- When using the tree browser in stand-alone mode, users reported seven time-outs, indicating that they could not solve a given task at all (in reasonable time), while only four time-outs occurred when using the combination of tree and telescope view. It is interesting to note that most users reported time-outs when they got completely lost in the

hierarchy and were unable to find a promising path of navigation towards a desired destination. Obviously, the telescope view is useful for comprehending the overall structure of the collection hierarchy and the current position in context.

- In both stand-alone views, about half of all time-outs occurred in task 2, which asked users to locate an item deep in the hierarchy by navigation. However, in mixed mode view only one user reported a time-out for task 2. This further underlines the importance of having both views to keep an overview.
- The labelling problems (i.e. "jumping" labels, overlaps, occlusion) reported in the initial evaluation were rarely mentioned by users in the interviews. Obviously, the strategy chosen to reduce these problems in the new prototype (merging labels when an overlap is likely to occur) is valid and conforms with user demands.

In general, continued development on the InfoSky visual explorer has yielded a much more stable and mature system. Many small usability issues observed by users in the initial evaluation did not come up in the recent study. Users appreciated the combination of tree and telescope view and, in interviews, consistently described having both available as more satisfying than any of the two alone. However, the problem remains that users have many more hours of training in conventional tree view interfaces than in the prototype telescope view.

We feel that with comparable amount of training and user experience the galaxy view would yield significantly better results than it is case the case now, and might – on occasions when context and overview can be exploited - outperform the tree view. On the other side, users at least partially familiar with the hierarchy will not profit from the galaxy view, and will probably be slower if offered only this view due to the additional cognitive load.

5 Related Work

Publications on the visualisation of large document repositories usually favorize either information retrieval based approaches utilising inter-document similarity measures within flat repositories, or visual exploration of hierarchically organised structures. Only recently have some first steps been taken towards integrating these two approaches.

5.1 Approaches Based on Inter-Document Similarity

Several systems employ methods for mapping documents from a high-dimensional term space to a lower dimensional display space, preserving the high-dimensional distances as far as possible in the process.

The Bead system [7] employs a thematic landscape view. The information space is arranged based on inter-document similarity forming a 2.1D landscape. Users can navigate freely around the information landscape. In contrast to InfoSky, Bead operates on flat document repositories and does not employ hierarchical structures.

Galaxy Of News [8] constructs and visualises associative relation networks 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 [9][10] operates on flat, unstructured document collections. Two visualisations are provided: 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. SPIRE's ThemeView (formerly Themescape) builds on the galaxy view by aggregating frequently occurring topical keywords from neighbouring documents and displaying the main themes in a thematic landscape. Documents matching particular search criteria can be grouped and colour-coded in the galaxy display.

Earlier work at the IICM on VisIslands [11][12] used standard clustering techniques to cluster document sets returned in response to a search query on the fly. The clusters were used for more efficient similarity placement, by first placing cluster centroids, and then placing documents around them.

WEBSOM [13] and other systems employ self-organising maps (SOMs) 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.

5.2 Approaches Based on Hierarchical Structure

Systems focusing on the visualisation and navigation of large hierarchical structures often optimise the use of available screen (pixel) real estate by geometric transformations and zooming and panning interactions.

The Hyperbolic Browser [14] is a two-dimensional tree browser, which utilises hyperbolic geometry to always display the entire hierarchy on the display. The H3 browser [15] 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 [16] lay out hierarchies in three dimensions. Each node in the hierarchy is the apex of a cone, with the root of the hierarchy being placed near the top of the three-dimensional display space and its children being evenly spaced along its base. 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. The shape of the visualisation is solely determined by the hierarchical structure, inter-document or intercollection similarities are ignored.

The File System Navigator (FSN) [17][18] uses a landscape metaphor to lay out a file system in three dimensions. Directories are represented as rectangular pedestals, successive subdirectories spread out in ranks back towards the horizon. Lines connecting the pedestals show the structure of the hierarchy and are traversable. Individual files are represented by boxes arranged atop each pedestal, the height of a box indicates the size of a file, while its colour represents its age. The layout, is determined purely by the structure of the hierarchy.

CyberGeo Maps [19][20] use a stars and galaxy metaphor to lay out pages of a web site. First, 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. While metaphor and visual display are similar to that used in InfoSky, the underlying layout is very different.

5.3 Integrated Approaches

Information Pyramids [21] 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. Children are arranged on top of their parents in a recursive fashion. 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 [22][23] 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.

6 Future Work

Work is continuing on the integration of the usability test results into the existing InfoSky implementation. A final version of the current InfoSky System is in development and will be put to extensive practical evaluation. Search functionality will be fully integrated with the tree and galaxy browser, and a separate usability study is designed to explore the power the combination of these components offers to users.

A visual classification algorithm will utilise the galaxy visualisation to display areas where new documents fit best, based on their content, and to allow users to directly (i.e. using mouse drag-and-drop) insert new documents into the hierarchy.

7 Concluding Remarks

We have presented InfoSky, a system for visual exploration of very large, hierarchically structured document repositories. After an initial user test designed to establish a comparison base line for further experiments, the system has been revised and extended to match user demands. A recent, more complex evaluation showed clear improvements over the initial prototype.

While several problems remain to be solved, using the telescope metaphor in conjunction with a conventional tree view displays clear benefits and justifies further development and evaluation of the InfoSky system.

We would like to thank our colleagues at the Know-Center, Hyperwave, and Graz University of Technology for their feedback and suggestions. The Know-Center is a Competence Center funded within the Austrian K plus Competence Centers Program (www.kplus.at) under the auspices of the Austrian Ministry of Transport, Innovation and Technology.

References

[1] Hyperwave R&D, patent holders Frank Kappe (Hyperwave R&D), Vedran Sabol (Know-Center) and Wolfgang Kienreich (Know-Center), 2002. Euopean and US patent pending.

[2] Ben Bederson. Jazz, 2002. http://www.cs.umd.edu/hcil/jazz/.

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

[4] Keith Andrews, Wolfgang Kienreich, Vedran Sabol, Jutta Becker, Georg Droschl, Frank Kappe, Michael Granitzer, Peter Auer and Klaus Tochtermann. The InfoSky Visual Explorer: Exploiting Hierarchical Structure and Document Similarities. In *Palgrave Journal on Information Visualisation, Issue 02/2002*, England, 2002.

[5] Matthew Chalmers. A linear iteration time layout algorithm for visualising high-dimensional data. In *Proc. Visualization'96*, pages 127–132, San Francisco, California, October 1996. IEEE Computer Society. http://www.dcs.gla.ac.uk/~matthew/papers/vis96.pdf.

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

[7] Matthew Chalmers. Using a landscape metaphor to represent a corpus of documents. In *Spatial Information Theory, Proc. COSIT'93*, pages 377–390, Boston, Massachusetts, September 1993. Springer LNCS 716. http://www.dcs.gla.ac.uk/~matthew/papers/ecsit93.pdf.

[8] Earl Rennison. Galaxy of news: An approach to visualizing and understanding expansive news landscapes. In Proc. UIST'94, pages 3–12, Marina del Rey, California, November 1994. ACM. http://www.acm.org/pubs/citations/proceedings/uist/192426/p3-rennison/.

[9] 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, 2001.

[10] James A. Wise. The ecological approach to text visualization. *Journal of the American Society for Information Science*, 50(9):814–835, July 1999. http://www.vistg.net/hat/Wise_draft/Ch5/Wise.html.

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

[12] Vedran Sabol. Visualisation islands: Interactive visualisation and clustering of search result sets. Master's thesis, Graz University of Technology, Austria, October 2001. ftp://ftp.iicm.edu/pub/theses/vsabol.pdf.

[13] Websom - self-organizing maps for internet exploration. Helsinki University of Technology, 2000. http://websom.hut.fi/websom/.

[14] John Lamping, Ramana Rao, and Peter Pirolli. A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In *Proc. CHI'95*, pages 401–408, Denver, Colorado, May 1995. ACM. http://www.acm.org/sigchi/chi95/Electronic/documnts/papers/jl_bdy.htm.

[15] Tamara Munzner. H3: Laying out large directed graphs in 3d hyperbolic space. In *Proc. IEEE InfoVis'97*, pages 2–10, Phoenix, Arizona, October 1997. IEEE Computer Society. http://graphics.stanford.edu/papers/h3/.

[16] 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.

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

[18] Joel D. Tesler and Steven L. Strasnick. Fsn: The 3d file system navigator. Silicon Graphics, Inc., 1992. ftp://ftp.sgi.com/sgi/fsn. http://www.jucs.org/jucs_7_6/discovering_knowledge_through_visual.

[19] Tobias Skog and Lars Erik Holmquist. Continuous visualization of web site activity in a public place. In *Student Poster, CHI 2000 Extended Abstracts*, The Hague, The Netherlands, April 2000. http://www.viktoria.informatik.gu.se/groups/play/publications/2000/WebAware.pdf.

[20] Lars Erik Holmquist, Henrik Fagrell, and Roberto Busso. Navigating cyberspace with cybergeo maps. In *Proc. of Information Systems Research Seminar in Scandinavia (IRIS 21)*, Saeby, Denmark, August 1998. http://www.viktoria.informatik.gu.se/groups/play/publications/1998/navigating.pdf.

[21] Keith Andrews, Josef Wolte, and Michael Pichler. Information pyramids: A new approach to visualising large hierarchies. In *IEEE Visualization'97, Late Breaking Hot Topics Proc*., pages 49–52, Phoenix, Arizona, October 1997. ftp://ftp.iicm.edu/pub/papers/vis97.pdf.

[22] WebMap. WebMap, 2002. http://www.webmap.com/

[23] Michael Iron, Roi Neustedt, and Ohad Ranen. Method of graphically presenting network information. US Patent Application 20010035885A1 , WebMap, November 2001. Filed 20th March 2001.