Abstract

It is no longer unusual for large document collections to contain many millions of documents. In order to manage this size of repository, it is often essential to structure the repository according to a thematic classification hierarchy.

InfoSky is a system enabling users to explore such large, hierarchically structured document collections. Similar to a real-world telescope, InfoSky employs a planar graphical representation with variable magnification. The hierarchical structure is represented by recursive Voronoi subdivision of the available space. At each level, larger subcollections are assigned more space and related subcollections gravitate towards one another.

Individual documents at each level of the hierarchy are represented by stars placed according to their similarity. Finally, InfoSky extracts and displays descriptors describing the essential theme of each subcollection at each level of the hierarchy.

Introduction

The sheer amount of information being managed and stored on computers ranging from laptops to network servers for global organisations is mind-boggling. The amount of new information generated in 2002 is estimated to be around 5 exabytes (1 exabyte = 10^{18} bytes), of which just over 90,000 terabytes is accessible through the web [6]. The world’s largest database in 2002 was at the Stanford Linear Accelerator Center, which stores 500 terabytes of experimental data.

Even though textual information only accounts for a fraction of the byte count, it is no longer unusual for large document collections to contain many millions of documents. In order to manage this size of repository, it is often essential to structure the repository in some way, often into a hierarchy according to a thematic classification scheme.

Systems such as the Hyperbolic Browser [5] and Information Pyramids [2] visualise large hierarchical structures while optimising the use of screen real estate, but make no explicit use of document content and subcollection similarities. Other systems such as Bead [3] and SPIRE [9] visualise large document collections, but operate on flat document repositories and do not take advantage of hierarchical structure.
Figure 1: The original prototype of the InfoSky explorer. The dataset shown consists of approximately 100,000 German language news articles manually classified into around 9,000 thematic collections and subcollections up to 15 levels deep.

**InfoSky**

InfoSky takes advantage of both hierarchical structure and document content. The hierarchy is assumed to be a topical or thematic classification hierarchy of collections and subcollections. Other organisations such as conference papers organised chronologically by year and conference, rather than topically, are possible but do not produce such thematically cohesive visualisations. The documents are assumed to have significant textual content, which can be extracted if necessary with specialised tools.

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.

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.

**InfoSky Galaxy**

The InfoSky galaxy maps the topical hierarchy to a recursive Voronoi subdivision of the display space:

- First, subcollection centroids are calculated bottom-up for each collection in the hierarchy by forming and averaging term vectors for each member (document or subcollection) of a collection.
Next, in top-down fashion, subcollection centroids are positioned within their parent’s allotted space (a polygon of some kind) according to their similarity with each other and neighbouring collections using a variation of Chalmers’ force-directed placement algorithm [4, 7].

A polygonal area is allocated around each subcollection centroid according to that subcollection’s “size” using modified, weighted Voronoi diagrams [8, pg. 128].

The overall result is a recursive spider’s web like subdivision of the display space for each level in the thematic hierarchy.

**User Testing**

A small formal experiment with 8 users in a counterbalanced design was run in October 2002 to establish a baseline comparison between the prototype InfoSky galaxy browser and a traditional tree browser. This study found that the tree browser performed better than the galaxy browser when they are used in isolation, mainly due to the familiarity of users with traditional explorer-like tree browsers.

Since then, the InfoSky galaxy browser has undergone several improvements. The current version is shown in Figure 2. We are currently (March 2004) testing the new InfoSky tree browser, galaxy browser, and synchronised combination of the two, against each other for a variety of browsing tasks.

**Concluding Remarks**

As development proceeds further, we believe that InfoSky will constitute an important step towards practical, user-oriented, visual exploration of large, hierarchically structured document repositories. Readers are referred to detailed descriptions of the first InfoSky prototype and user study in [1].
References


