Challenges in Knowledge Discovery: Structured Repositories and Multimedia Content

Wolfgang Kienreich
(KnowCenter Graz,
wkien@know-center.at)

Vedran Sabol
(KnowCenter Graz,
vsabol@know-center.at)

Michael Granitzer
(KnowCenter Graz,
mgrani@know-center.at)

Mattias Lux
(KnowCenter Graz,
mlux@know-center.at)

Werner Klieber
(KnowCenter Graz,
wklieber@know-center.at)

Walter Sarka
(KnowCenter Graz,
wsarka@know-center.at)

Abstract. Recent trends in structure and content of global knowledge spaces present new challenges to the field of Knowledge Discovery. Very large, highly structured repositories are increasingly replacing smaller, flat information spaces. Such repositories are often filled with multimedia documents, including image, audio and video data. This publication briefly outlines the underlying trends and discusses implications on approaches to Knowledge Discovery. Some examples for applications accomodating these implications are presented and analysed for lessons learned which can be incorporated in designing future Knowledge Discovery systems. Emphasis is given to the visualisation of hierarchical structures and to cross-media knowledge mining, two fields crucial for adressing future challenges to Knowledge Discovery.

Key Words: Knowledge Discovery, Cross-Media Knowledge Mining, Structural Retrieval, Large Hierarichal Repositories, Information Visualisation, MPEG7, Metadata

Categories: H3.4, H3.7, H 5.1, H5.2

1. Introduction

A commonly used definition of Knowledge Discovery is that it describes the overall process of finding and identifying data of interest, extracting and generating knowledge from raw data, and presenting the results in an adequate and convenient way. Specific to Knowledge Discovery in Databases, an area of Knowledge Discovery which is often used synonymously for the whole field, another prevalent definition is that of *the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data* [Fayad, 96]. Recent trends in amount and type of information available on the worldwide web as well as in corporate databases and intranets introduce upcoming challenges for the field as a whole, and for the above-given definitions in especial. We will focus on two particular developments, the increase in multimedial content available in the commercial and consumer area and the advent of very large, highly structured knowledge repositories in corporate intranets. Both developments will be characterized briefly, and examples for systems adressing them will be provided.

1.1 Multimedia content

The advent of digital cameras as well as the massive sharing of digital audio files through dedicated peer-to-peer software has lead to an explosion in the number of multimedia documents available. It is estimated that there will be nearly 300 million digital image capture devices in use worldwide through 2004, capturing about 29 billion digital pictures, most of which will be organised in some kind of multimedia repository and available via the world wide web or other means of sharing data [Infotrend, 04].

Classical Knowledge Discovery had a strong focus on Data Mining as *a blend of statistics, artificial intelligence, and data base research* [Pregibon, 97], aiming at the extraction of trends and patterns from numerical or text databases [Kodratoff, 99]. Extending this approach to multimedia archives is a non-trivial task: Multimedia information includes richer and more diverse content than text-based information so that concepts like “relevance” or “similarity”, which are central to statistical and artificial intelligence methods, become much more difficult to model and capture.

To encompass multimedia knowledge repositories, Knowledge Discovery has to expand its range of retrieval and visualisation concepts to areas like content-based image retrieval, retrieval of annotated meta data and multimedia document visualisation. The vision of cross-media Knowledge Mining, the process of integrating images, texts, numerical data, audio data and other formats into a single consistent representation which can be queried in a unified way, is a key topic for future activities in Knowledge Discovery.

1.2 Large, structured repositories

Despite a less-than-bright economic situation, worldwide spendings on knowledge services are expected to increase by over 40% per year, and a major part of that spendings will be invested in the area of Knowledge Management [Sandhya, 02]. Contrary to classical means of storing corporate information, modern Knowledge Management Systems organize knowledge in a highly structured and interlinked way. For example, the Hyperwave eKnowledge Suite [Hyperwave, 04] which is based on the HyperG technology [Maurer, 96] organises knowledge objects in the form of a directed graph and presents the user with a tree structure for browsing the knowledge space. Other products like Lotus or OpenText incorporate similar functionality.

With an increasing number of companies organising their Intranets based on such systems, classical databases or flat repositories of documents are becoming a thing of the past. The typical corporate knowledge space of the very near future is shaped into a hierarchy or some other, probably more complex kind of graph, with the structural information contained in this organisation comprising an integral and vital component of the overall knowledge contained.

Initiatives in Semantic Web technologies are trying to structure the World Wide Web as a whole in a similar manner. However, even if the vision of a global distributed knowledge space maintained on a peer-to-peer basis does not become reality in the midterm, Semantic Web technologies provide a huge longterm potential in application domains like Enterprise Information Discovery and Integration, E-Commerce and general Knowledge Management [Tolksdorf, 03]. Consequentially, Knowledge Discovery systems must be aware of structural information and exploit it in all aspects of their operation and functionality. Structural retrieval and visualisation of structured repositories are key issues in future activities in Knowledge Discovery.

2. Magick: Knowledge Discovery in Multimedia Data

Locating information on the world wide web, in the sense of classical web content mining, usually includes the extraction of textual content and page elements in the form of (html) tags and the computation of correlation between extracted information to gain new insights [Cooley, 99]. With the increasing amount of digital images presented on the word wide web and the often sparse annotation available, efforts to combine correlation or similarity measures obtained from different formats into a single consistent space are gaining importance. In the Magick project, clustering and multidimensional scaling have been applied to a set of data containing text and image documents of various formats comprising a space of semantically related objects where image documents outnumber textual content by a large margin. As an example, imagine a web page about "medieval castles" which contains some describing text and several photos and outlines of prominent castles. Splitting up this web page into one or more text documents, a number of images and some annotations or meta data entries yields an object space containing objects of differing formats which are nevertheless semantically related. Gathering a number of similar web pages (i.e. by "spidering" through link following) would quickly build up a multimedia object repository challenging most existing Knowledge Discovery systems.

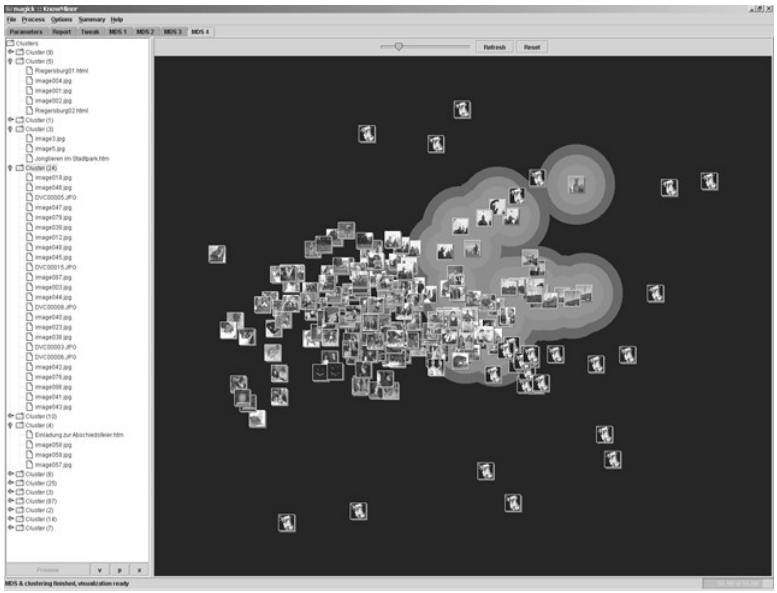


Figure 1: The Magick user interface

Magick employs automated clustering and multidimensional scaling of all objects in a repository to generate a user interface featuring a tree view of clusters (Figure 1, left) synchronized with a two-dimensional map view (Figure 1, right) of clusters and documents. A central requirement of the clustering and scaling techniques employed is the ability to compute the similarity between two objects in the semantic repository. Magic achieves this by using a variety of different metrics termed feature spaces, depending on object type, which can be weighted to combine into a global similarity measure.

The similarity measurement between single documents is based on the document metadata. The metadata elements available for each document format are mapped to

Dublin Core metadata elements. For example, EXIF [JEITA, 02] creation date or file creation date for digital photos was mapped to the date element in Dublin Core. So for all documents Dublin Core based metadata was available regardless of media type. In addition to the Dublin Core feature space media-specific feature spaces were implemented and used for similarity computations between documents of the same media type. For example, and EXIF feature space for digital photos, an IPTC feature space for annotated images and two additional content based image retrieval feature space based on the MPEG-7 descriptors ScalableColor and ColorLayout [Kasutani, 01] were implemented. All the similarities calculated in different feature spaces were combined and weighted to create a parametrized overall similarity upon all feature spaces .

Lessons learned Preliminary evaluation of Magick demonstrates that while the option to have images, text and metadata processed within a consistent knowledge space, and to have all these elements presented within the same visualisation, is of great benefit for users in locating and filtering information of interest, the balance of the various measures remains an open problem. With so many intervening variables, it is often not clear for users why exactly a specific piece of information is assigned to cluster A instead of Cluster B. Interactive manipulation of the weights applied to the various measures aids users in understanding what is happening, but the underlying problem remains. We suppose that some kind of user and task specific profiling will be necessary.

3. InfoSky: Knowledge Discovery in very large, hierarchically structured repositories

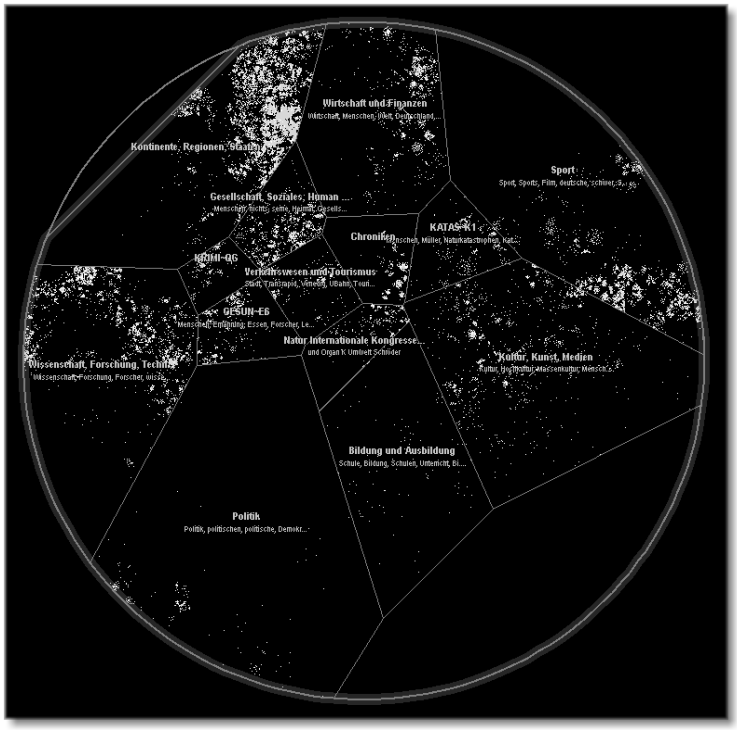
The InfoSky visual explorer is a system enabling users to interactively explore large, hierarchically structured document collections. Similar to a real-world telescope, InfoSky employs a planar graphical representation with variable magnification. Documents of similar content are placed close to each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

InfoSky features sophisticated search functionality, including the ability to execute a number of independent queries. Results of each query are displayed as colour-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.

InfoSky adresses the issue of large, hierarchical repository structures by exploiting the hierarchy for performance optimisation and by shaping the resulting visualisation towards reflecting the hierarchy. Only a brief overview of algorithmic details can be given in this publication, for an in-depth analysis see, for example, [Andrews, 02].

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 subcollections

Figure 2: The Infosky visualisation



across collection boundaries tend towards each other (they are not allowed to actually cross the boundary). Similarity placement is realised using an optimised force-directed placement algorithm [Chalmers, 96]. The layout in normalised 2D space is transformed to the polygonal area of the parent collection. 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 using a modified Voronoi diagram [Okabe, 00]. Finally, documents contained in the collection at this level are positioned using the similarity placement algorithm, according to their interdocument similarity and their similarity to the sub-collection centroids at this level, which are used as static influence factors. The final result is a visualisation which combines hierarchy (collection bounds) and documents (stars) into a single visual metaphore.

A number of available information visualisation systems provides a presentation similar to that of InfoSky. However, these systems do not incorporate hierarchical structure into their computation or visualisation algorithms. For example, SPIRE [Thomas, 01] provides an Information Galaxy as a visualisation, but operates on flat, unstructured document collections and does not exploit any inherent hierarchical structure. The Bead system [Chalmers, 93] employs a thematic landscape view where the information space is arranged based on inter-document similarity forming a 2.1D landscape similar to InfoSky's galaxy. Bead, too, operates on flat document repositories and does not employ hierarchical structures.

InfoSky has been under development and evaluation for more than three years. Still, recent user tests [Granitzer, 04] demonstrated that while the system displayed increased performance as compared to earlier prototypes, users were faster, by a decisive margin, for a range of tasks when using traditional tree browsers than when using the InfoSky visual explorer. However, a combination of the tree browser and InfoSky

yielded significantly better results than the use of InfoSky alone, and users got lost in deep hierarchies significantly less often with the additional help of the visualisation.

Lessons learned The InfoSky approach of exploiting hierarchical structure for performance optimisation instead of handling it as an additional obstacle has proven quite successful. Hierarchies containing millions of documents have been processed into a galaxy visualisation overnight (ten hours), clearly demonstrating the technical achievement. However, extensive user testing of the system revealed that users are still more comfortable with at least some traditional user interface elements present, and perform best when using such elements instead of more innovative ones tailored towards the tasks at hand. Consequentially, user interfaces should be designed conservatively, with few and task-oriented new components present, even if the underlying mining technology is advanced and offers lots of features.

4. Conclusions

In this paper, we have discussed some challenges to the field of Knowledge Discovery resulting from recent trends in the structure and content of emerging global knowledge spaces. We then briefly outlined some projects which dealt with designing and implementing Knowledge Discovery systems trying to meet two of these challenges, cross-media knowledge mining and discovery of knowledge in highly structured data. From our experience, we derived various lessons learned.

We found that cross-media knowledge mining provides results not immediately transparent to users. Interactivity or profiling should be employed to ease the cognitive load on users and to help them understand what, for example, a complex similarity metric actually expresses in real-world terms. We identified deep hierarchies as valuable resources for optimisation of visualisation algorithms, but determined that specialised user interfaces require lots of training for users accustomed to conventional means of search and navigation, so integration of advanced Knowledge Discovery components with standard user interfaces should be propagated to let a wide range of users profit.

5. Future Work

Based on our experience in the projects described in this paper as well as in other, related work, we find it possible to devise a framework encompassing the Knowledge Discovery process as a whole while adressing challenges like cross-media knowledge retrieval or highly structured knowledge spaces. We have recently set upon designing an implementing such a framework: In a project named KnowMiner, we are developing a cross-media Knowledge Retrieval and Visualisation framework.

The KnowMiner Framework will be an extendable, configurable, general-purpose-knowledge discovery software package implemented in a portable, platform-independent way and offering standardized, XML-based input/output capabilities. KnowMiner will focus on handling human-readable cross-media information, and act as a bridge between data repositories and front-end interfaces like browsers or rich clients. The framework will handle information available in different formats, stored in heterogeneous repositories, but also offer powerful means of analysis, orientation and navigation in large cross-media data repositories.

As a consequence of our experiences in the Magick project, KnowMiner will include extensive support for profiling and configuration of processing algorithms, to allow tuning of applications to users requirements and perception of results. Based on lessons learned from InfoSky, KnowMiner will not offer any built-in, fixed visualization or presentation means. However, its functionality, configurability and standardized interfaces will ensure that presentation layers can be implemented atop the framework with a minimum effort.

We hope that we will be able to complete the KnowMiner framework, to apply it to a number of real-world problems from the Domain of Knowledge Discovery, and to evaluate it extensively in the near future, in the face of the Challenges we have described in this publication.

6. Acknowledgements

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

[Kodratoff, 99] Y.Kodratoff, Knowledge Discovery in Texts: A Definition, and Applications in Foundation of Intelligent Systems, Ras&Skowron, LNAI 1609, Springer 1999.

[Global, 04] Global Reach Inc., Internet, 2004, http://www.global-reach.biz

[Nielsen, 04] Nielsen Netratings Inc, Internet, 2004, http://www.nielsen-netratings.com

[Marcussen, 04] C. Marucssen. Centre for Regional and Tourism Research, Denmark. Internet, 2004, http://www.crt.dk/uk/staff/chm/trends.htm

[Pregibon, 97] D.Pregibon: Data Mining in Statistical Computing and Graphics, American Statistical Association, 1997

[Cooley, 97] R. Cooley, B. Mobasher, J. Srivastava: Web Mining: Information and Pattern Discovery on the World Wide Web. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), November 1997

[Sandhya, 02] S.M.Sandhya: KM market: Eyeing exponential growth. CIOL, Internet, 2002, www.ciol.com

[Infotrend, 04] Infotrend Research Group, Inc., 2004, Internet, www.infotrends-rgi.com

[Maurer, 96] H. Maurer: Hyperwave - The Next Generation Web Solution. Addison Wesley: Harlow, UK. 1996.

[Hyperwave, 04] Hyperwave R&D, Internet, 2004, www.hyperwave.com

[Fayyad, 96] U.Fayyad, W.Piatetsky-Shapiro, S.Smyth: "From Data Mining to Knowledge Discovery: An Overview", Advances in Knowledge Discovery and Data Mining, AAAI Press/MIT Press, Menlo Park, CA, 1996, pp.1-34

[Kasutani, 01] E. Kasutani, A. Yamada: ”The MPEG-7 Color Layout Descriptor: A Compact Image Feature Description for High-Speed Image/Video Segment Retrieval”, Proc. of International Conference on Image Processing (ICIP 2001), vol. I, pp. 674-677, October 2001.

[Andrews, 02] K.Andrews, W.Kienreich, V.Sabol, J.Becker, G.Droschl, F.Kappe, M.Granitzer, K.Tochtermann, P.Auer: The InfoSky Visual Explorer: Exploiting Hierarchical Structure and Document Similarities, Palgrave Macmillan, 2002, 166-181

[Granitzer, 04] M.Granitzer, W.Kienreich, V.Sabol, K.Andrews, W.Klieber: Evaluating a System for Interactive Exploration of Large, Hierarchically Structured Document Repositories. Submittet to InfoVis 2004,tenth annual IEEE Symposium on Information Visualization, Austin , Texas, 2004

[Thomas, 01] J.Thomas, P.Cowley, O.Kuchar, L.Nowell, J.Thomson, P. Chung Wong. Discovering knowledge through visual analysis. Journal of Universal Computer Science, 7(6):517–529, 2001.

[Chalmers, 93] M.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

[Tolksdorf, 03] R.Tolksdorf, C.Bizer, R.Eckstein, R.Heese: Business to Consumer Markets on the Semantic Web. OTM Confederated International Workshops, HCI-SWWA, IPW, JTRES, WORM, WMS, and WRSM 2003, Catania, Sicily, Italy, 2003.

[JEITA, 02] Japan Electronics and Information Technology Industries Association (JEITA): Exchangeable image file format for digital still cameras - Exif Version 2.2, 2002.

[Chalmers, 96] M.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

[Okabe, 00] Atsuyuki Okabe, Barry Boots, Kokichi Sugihara, and Sung Nok Chiu. Spatial Tesselations: Concepts and Applications of Voronoi Diagrams. Wiley, second edition, 2000. ISBN 0471986356.