CALL FOR PAPERS

1st Workshop on

Knowledge Management for Distributed Agile Processes: Models, Techniques, and Infrastructure (KMDAP '03)

for the

Twelfth IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET ICE '03).

June 9-11, 2003, University of Linz, Linz, Austria

Up-to-date information on the workshop can be found at

http://www.dwm.uni-hildesheim.de/homes/schaaf/WETICE03/

Workshop Description

Knowledge Management (KM) is currently receiving increasing attention in diverse areas such as medicine and systems engineering. Here, special focus is put on process-oriented Knowledge Management, where abstract activity descriptions serve as the primary means to capture, organize, and distribute knowledge items that are relevant during individual, actual process steps. Most approaches developed so far rely on static processes as well as on documents indexed by formalized meta-data and additional ontologies. However, these approaches are inadequate for highly dynamic and volatile processes, whose steps cannot be planned in advance, and during which new, unanticipated "knowledge needs" frequently arise. Such processes handle mostly informal documents and rely on face-to-face communication between participants. Typical examples of such processes occur in domains like medical diagnostics and disaster management.

In Software Engineering, the realization that software development processes are inherently dynamic inspired an entire new discipline focusing on "Agile Software Development Processes". These human-centered methodologies are being increasingly applied in the past couple of years. However, trading off explicit knowledge captured in documentation for tacit interpersonal knowledge poses new challenges, especially in the case of distributed settings, where support by proper Knowledge Management techniques is essential.

The main goals of this workshop are to bring together practitioners and researchers from the areas of Knowledge Management and Agile Processes from different domains to discuss the current state of ongoing research efforts and to share their practical experiences with adaptation of modern Knowledge Management techniques by agile teams.

Topics of Interest

Topics of interest include, but are not limited to, the following areas:

- Knowledge management support systems for agile teams
- How to remain agile while applying KM techniques relying on explicit knowledge representation?
- KM techniques that help making processes more agile
- KM to improve agile processes (self-adaptive processes)

- Collaborative ontology construction and mediation
- Knowledge assets of agile teams
- Knowledge elicitation in distributed agile processes
- Proactive knowledge distribution
- Cooperative adaptation of knowledge in agile processes
- Managing tacit knowledge
- Inter-project knowledge integration and management
- Knowledge visualization for agile teams
- Knowledge annealing

In the context of this workshop, our notion of distribution is a wide one, encompassing any situation where direct face-to-face communication between (current or former) process participants is somehow inhibited.

Concerning the notion of agility, a number of interpretations have been developed by different industries (e.g., see [2] for the Software Industry). For the purpose of this workshop, we adopt the encompassing view from [1], where business agility is defined as "the ability to demonstrate flexible, efficient and swift responses to changing circumstances by maximizing physical and human resources."

--

[1] Gartner UK Ltd.: "The Age of Agility", Report prepared by Gartner for BT, July 2002

[2] http://www.agilemanifesto.org

Paper Submission

Papers (maximum: 6 pages in 10-pt Times, single-spaced) can be submitted for review in PDF or RTF format. Papers longer than 6 pages will not be reviewed.

Please submit your papers via e-mail to

Harald Holz
e-mail: holz@informatik.uni-kl.de

Authors will be requested to take part in the peer review process and will be asked to review other submissions. In order to ensure an anonymous review process, please try to avoid including any information in the body of the paper or references that would identify the authors or their institutions. Instead, please provide the names and contact information of the authors in your submission e-mail. This information can also be added to the final camera-ready version for publication at a later stage.

Publication

Accepted papers will be published in the post-conference proceedings (publisher: the IEEE Computer Press). Final camera-ready copies may not exceed six pages and must conform to the IEEE Computer Society Press Proceedings Author Guidelines, http://www.computer.org/cspress/instruct.htm

Each accepted paper should have at least one author register and present the paper at WETICE-2003 to get the paper published in the Proceedings.

Important Dates

Paper submission deadline:	March 7, 2003
Notification of acceptance:	April 11, 2003
Final camera ready copies of accepted papers due to IEEE:	May 16, 2003
Advance registration by:	May 16, 2003
Workshop:	June 9-11, 2003

Workshop Location

Linz is the Upper Austrian state capital, third largest city in Austria, and located on both sides of the Danube. For more information on Linz, please see:

http://www.hotelreservationandinformation.com/cityguide.phtml?city=Linz&stateprovince=&country=Austria

For more information on Austria, please see:
http://www.austria-tourism.at/index_e.html

Organizing Committee

Harald Holz, University of Kaiserslautern, Germany
Grigori Melnik, University of Calgary, Canada
Martin Schaaf, University of Hildesheim, Germany

Program Committee

Klaus-Dieter Althoff, Fraunhofer Institute for Experimental Software Engineering (IESE), Germany
Ralph Bergmann, University of Hildesheim, Germany
Ansgar Bernardi, German Research Center for Artificial Intelligence (DFKI) GmbH, Germany
Ward Cunningham, Cunningham & Cunningham, Inc., USA
Mehmet Goeker, Kaidara Software Inc., USA
Scott Henninger, University of Nebraska-Lincoln, USA
Frank Maurer, University of Calgary, Canada
Charles Petrie, Stanford University, USA
Michael M. Richter, University of Kaiserslautern, Germany
Steffen Staab, University of Karlsruhe (TH), Germany
Laurie Williams, North Carolina State University, USA

Please contact Harald Holz (holz@informatik.uni-kl.de) with questions and/or suggestions.

WebRat: Supporting Agile Knowledge Retrieval through Dynamic, Incremental Clustering and Automatic Labelling of Web Search Result Sets

Michael Granitzer
KnowCenter Graz
mgrani@know-center.at

Vedran Sabol
KnowCenter Graz
vsabol@know-center.at

Wolfgang Kienreich
KnowCenter Graz
wkien@know-center.at

Abstract

WebRat is an interactive system for visualising and refining search result sets. Documents matching a query are dynamically clustered on the fly and visualised as a contour map of islands. Thematic clusters are built, analysed, and visualised in real time.Users can interactively explore the visualisation and refine queries by selecting from the keywords and clusters presented to them. WebRat does not rely on precalculated meta data. Instead, necessary information is directly extracted from query result representations provided by search engines, as for example ranked lists of document snippets. The system is language-independent, does not require a dedicated server machine and can be adapted to a number of data sources and visualisation modes easily. WebRat supports agile knowledge retrieval by transforming unstructured information input into a representation enriched with structure and meta information.

1. Introduction

Todays standard web search interfaces are very much alike in user interface and implementation detail: Users type in one or more textual query terms and are then presented with a ranked list of matching documents in decreasing order of relevance, based on a full-text search of the query terms in a given data set. While easy to implement and maintain, such an approach features a number of drawbacks which renders it less usefull in a complex query context featuring, for example, multimedia content or relationships between objects.

On the user interface side, a ranked list can only display a small subset of results if the result set is large, and the manifold topical dimensions of the result are hidden from the user. On the query and retrieval side, full text engines are often tailored towards a specific datasource (i.e., a database of given structure) and do not easily adapt to varying datasources or differently structured types of information. Hence, knowledge workers get suboptimal search solutions, and software engineers have to reinvent the wheel every time they create a search solution for a new domain of knowledge.

Recently, many proposals have tried to adress these issues by enriching unstructured repositories with meta-data, or by introducing structures like topic maps and ontologies to represent topical interconnections and, in general, support search operations [CITE: TOPICMAPS]. While such approaches work well in clearly specified areas like the environmental domain, where rich meta-data is already available, they fail in situations where annotation or structuring of information entities is complicated or not possible at all. [CITE: ENVIRONINFO].

The WebRat retrieval and visualisation system was designed to address these problems. WebRat provides an agile framework capable of:

- querying various web data sources (in the fashion of a metasearch engine);
- dynamic, incremental clustering of search results by topic;
- automatically extracting keywords describing topics and using these as cluster labels;
- interactive visualisation of results and topics in a number of ways.

The system does not require any precalculated information., as all necessary operations are done on the fly, based on search results as they arrive. All calculations can be performed on standard office machines, visualisation works with low-performance graphics hardware, and no dedicated server or service environment is necessary. We believe that WebRat's innovative features and total independence of the presence of any static information will provide a more satisfactory search and retrieval experience to knowledge workers and web users.

2. WebRat

2.1. Overview

WebRat supports the identification of topical groups of information entities within search results through dynamic, incremental clustering. A thematic landscape of matching documents is generated and updated on-the-fly as search results arrive. Query refinement is simplified by labelling thematic clusters with automatically extracted keywords. Users start the query process by entering a number of query terms. These terms are sent to all of the data sources (i.e. search engines, databases, archives) selected by the user. The returned results are processed by the system and displayed in the visualisation.

The visualisation is updated at regular intervals as more and more results arrive. The results form islands on a virtual contour map which are labelled with the according keywords. User interaction includes zooming in and out to reveal more details or displaying an overview as well as navigating from island to island. Labels are calculated on the fly, so as to always describe the most obvious concentrations of documents (the largest islands) in the visualisation.

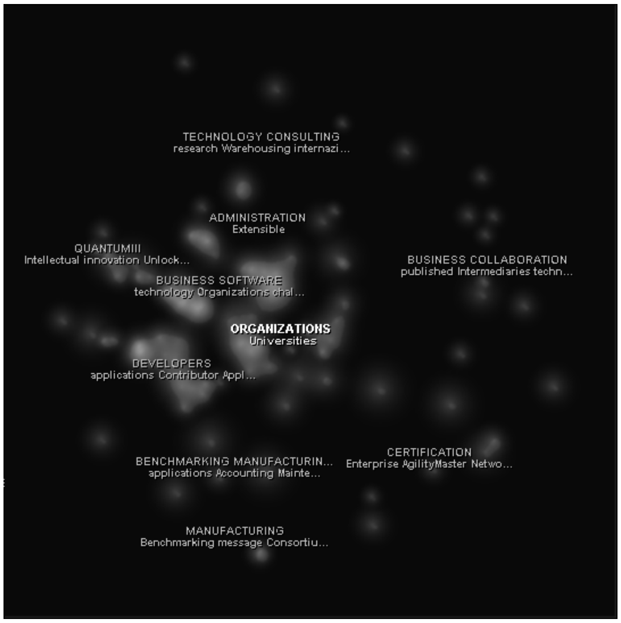
TECHNOLOGY CONSULTING
research Warehousing internazi...
ADMINISTRATION
Extensible
QUANTUMIII
Intellectual innovation Unlock...
BUSINESS COLLABORATION
published Intermediaries techn...
BUSINESS SOFTWARE
technology Organizations chal...
ORGANIZATIONS
Universities
DEVELOPERS
applications Contributor Appl...
CERTIFICATION
Enterprise AgilityMaster Netwo...
BENCHMARKING MANUFACTURIN...
applications Accounting Mainte...
MANUFACTURING
Benchmarking message Consortiu...

Figure 1: WebRat displaying the results of a query for “agile knowledge management”

Users can call up a context menu for each label which allows to re-queryany of the used data sources by refining the original query with the label keywords. In this way, the system supports the user in formulating a narrower search query. Additionally, the context menu contains commandsfor the basic navigational functionality described above. Finally, abstracts extracted from query results can be displayed for clusters as well as for individual documents, providing further decision support for a user faced with very similar results.

WebRat provides the described functionality without depending on any information but the document snippets and relevance data returned by a web search engine as a query result. A typical retrieval process using WebRat might consist of the following steps:

- The user enters some initial query terms, which probably are very general in nature.
- The user picks topics of interest from the visualised result set.
- The user launches a new search through a topic of choice, refining the initial search.
- The user browses the refined result set from the search engine(s) specified.

Figure 1 shows a typical WebRat visualisation. A query on “agile knowledge management” has been sent to two web search engines, and results have been organised into an island landscape, with each labelled island representing an outstanding topic within the returned result set. In the actual application, the user is able to watch the landscape grow and evolve as more and more results arrive.

The system appears to be very useful in cases when a user is dealing with an unfamiliar knowledge domain. Through the ability to quickly and dynamically structure the retrieved documents an overview and a point of entry for the domain is created. In such cases the user can learn about different subtopics the knowledge domain is composed of, and about terminology which is commonly used. Clicking on the labels of interest and following the more detailed subtopic descriptions helps understanding the internal structure of the knowledge domain, learning the significance of particular subtopics comprehending the relationship between subtopics. This aspect makes the WebRat particularly suitable for agile knowledge retrieval, where users are confronted with new or redefined requirements and/or information sources frequently.

2.2. Algorithms and Architecture

While a detailed discussion of the algorithms used in WebRat is beyond the scope of this paper (refer to [CITE: PAKM] for more information), the general architecture of the framework and several prominent algorithms can be briefly outlined. The WebRat framework consists of four components, which interact through a messaging system and a shared pool of objects. The components are:

High-Dimensional Component: Retrieves the search results from the search engines and creates a high-dimensional representation for each result. A *Grabber Unit* retrieves search results from a variety of data sources and adds them to a global document pool. Then, a *Vectoriser Unit* analyses document snippets employing a language-independent method known as n-gram decomposition. [CITE:NGRAMS] Subsequently a TF-IDF scheme is applied to the resulting high-dimensional representation of the retrieved documents [CITE:TFIDF]. The *High-Dimensional Centroid Computation Unit* computes the high-dimensional centroids for new clusters and continuously updates the existing clusters as new documents are inserted. Finally, a *Keyword Extractor* applies a weighting scheme to identify terms (n-grams) which best describe clusters and/or regions of interest in the 2D layout. Subsequently keywords are extracted by identifying the words and text segments which were the sources of the n-grams with the highest rating.

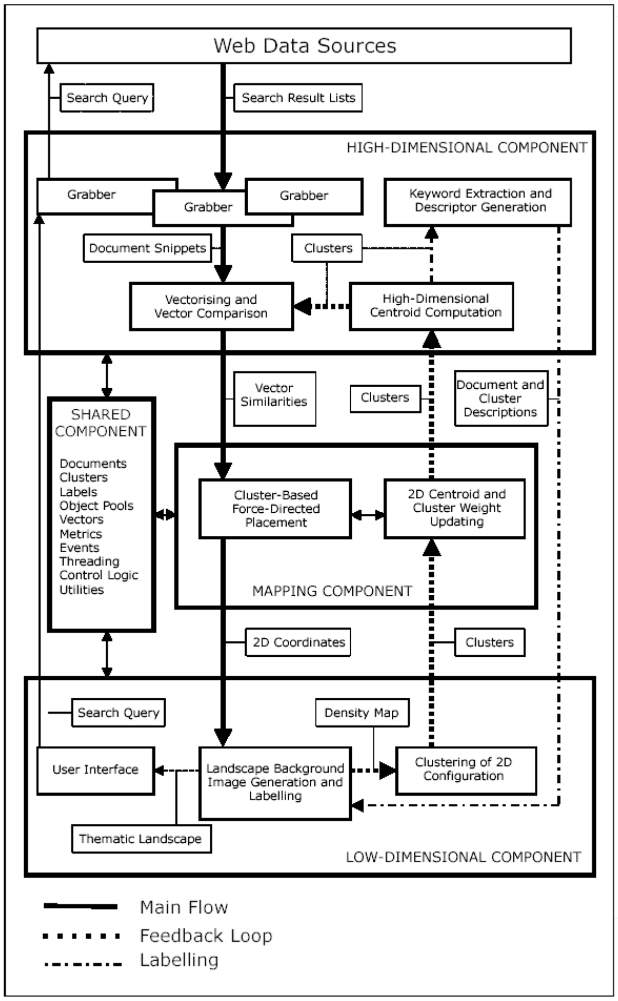
Web Data Sources
Search Query
Search Result Lists
HIGH-DIMENSIONAL COMPONENT
Grabber
Grabber
Grabber
Keyword Extraction and Descriptor Generation
Document Snippets
Clusters
Vectorising and Vector Comparison
High-Dimensional Centroid Computation
Vector Similarities
Clusters
Document and Cluster Descriptions
SHARED COMPONENT
Documents
Clusters
Labels
Object Pools
Vectors
Metrics
Events
Threading
Control Logic
Utilities
Cluster-Based Force-Directed Placement
2D Centroid and Cluster Weight Updating
MAPPING COMPONENT
2D Coordinates
Clusters
Search Query
Density Map
User Interface
Landscape Background Image Generation and Labelling
Clustering of 2D Configuration
Thematic Landscape
LOW-DIMENSIONAL COMPONENT
Main Flow
Feedback Loop
Labelling

Figure 2: WebRat System Architecture

Mapping Component: Performs the dimensionality reduction from the high-dimensional term-vectors to the 2D visualisation space. A *Force-Directed Placement (FDP) Unit* performs multidimensional scaling of the data set: based on high-dimensional term vector similarities. [CITE: FDP] 2D document positions are computed preserving the high-dimensional relations as far as possible. The FDP algorithm can operate in a cluster-oriented mode to improve performance and layout separation. In this mode the system tracks the creation of clusters in the layout and the the *Low-Dimensional Centroid Updater* continuously recomputes the low-dimensional cluster centroid positions as the clusters' children are repositioned by the layout algorithm.

Low-Dimensional Component: Consists of a user interface, a landscape generator and a 2D clustering module with feedback to the highdimensional component. The *User Interface* presents the visualisation and handles the interactivity and navigation. The *Landscape Generator* computes a shaded islands landscape based on the 2D document density, employing some innovative optimisation techniques to allow visualisation on standard office machines. The *2D Clustering Unit* analyses the landscape to identify positions of the 2D document density maxima created by the FDP algorithm. These are used as cluster seeds. Clusters are then created by assigning each document to the nearest seed. In doing so, a feedback loop is created to the FDP algorithm, which uses the created clusters to improve performance and layout quality.

Shared Component: Consists of document and cluster pools, high-dimensional and low-dimensional metric definitions, threading and control logic facilities, an event model, as well other components shared by the three processing components.

For maximum adaptability to various data sources and visualisation metaphors, each component is separately configurable and exchangeable. The components operate as Java threads, communicate with each other through a shared data pool, and use an event model for exchanging messages or requesting specific operations.

3. Applications

While initially the focus of WebRat has been on retrieval and visualisation of web search engine results, the framework has in the meantime been applied to a number of domains featuring various datasources and visualisation needs.

3.1. Web Query Refinement

The first and obivious use of the WebRat framework is its ability to run web metasearches. While there is a number of meta search engines available, none of those features the combination of clustering and visualisation capabilites WebRat provides. Consequentially, there is a WebRat web search engine page available at the homepage of the KnowCenter since November 2002 (http://www.know-center.at/webrat). Users can enter query terms, view results in a Web browser., refine their searches and experiment with a number of different visualisations.

The WebRat search engine has already undergone a first sample usability test where co-workers of the Know-Center known to do research-intensive work were asked to use their search engine of choice in parallel with WebRat for some time. Preliminary evaluation results indicate that users liked the ability to get an overview on the central topics (keywords) of an area new to them.

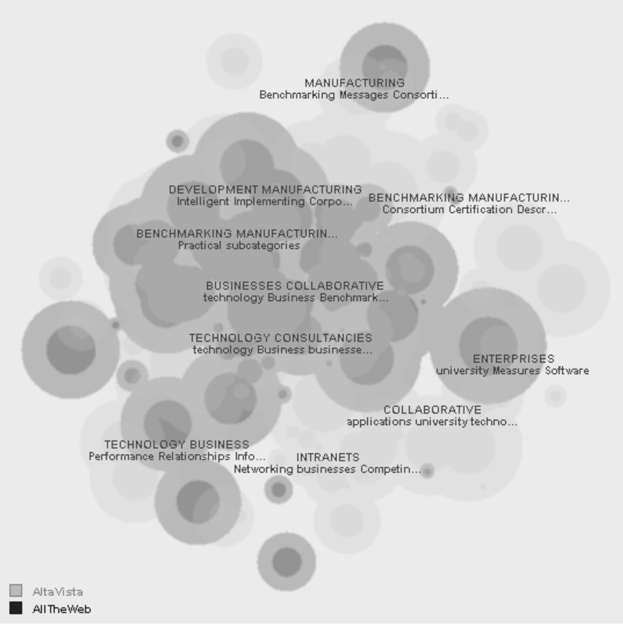
MANUFACTURING
Benchmarking Messages Consorti...
DEVELOPMENT MANUFACTURING
Intelligent Implementing Corpo...
BENCHMARKING MANUFACTURIN...
Consortium Certification Descr...
BENCHMARKING MANUFACTURIN...
Practical subcategories
BUSINESSES COLLABORATIVE
technology Business Benchmark...
TECHNOLOGY CONSULTANCIES
technology Business businesse...
ENTERPRISES
university Measures Software
COLLABORATIVE
applications university techno...
TECHNOLOGY BUSINESS
Performance Relationships Info...
INTRANETS
Networking businesses Competin...
AltaVista
AllTheWeb

Figure 3: Comparing Search Result Sets with WebRat

3.2. Qualitative comparison of search result sets

When multiple data sources are available for querying, it is often interesting to know how the result sets returned compare to each other. WebRat has been adapted to allow users to compare data sources using a visualisation where each entry returned is colour-coded according to its source. Figure 3 displays how two well-known search engines compare to each other with a query for the term “agile Knowledge Management”.

While this variant of WebRat does not give any scientifically sound measure for data source quality (which complex precision and recall analysis could provide), it is nevertheless a valuable qualitative tool for users which helps to determine strengths and weaknesses of data sources. The above image, for example, clearly shows that the green (light-grey) source has found a cluster on “Collaboration” which is invisible to the red (dark-gray) source.

3.3. Content-Based Query refinement for the environmental domain

The characteristics of environmental information make it a challenging field for search engines and query refinement tools. Environmental information is often made up of a variety of different data types, and is typically enriched with meta-information. The environmental context is saturated with abbreviations and multiple meanings of words, rendering the snippet information returned by standard web queries mostly useless. In this context, WebRat has been applied as a retrieval tool for querying two German environmental data catalogues. Meta-information returned by these systems was incorporated and given priority compared to snippet information. In addition, the visualisation was thematically tailored to the domain to demonstrate the adaptiveness of the WebRat

Comparisson of WebRat with the strongly meta-data based Paddle system [CITE: ENVIRONINFO] has shown that Paddle’s result set quality is highly dependent on amount and quality of available meta-data, and that Paddle isn’t well suited to providing an overview over a large, unspecific result set returned by a “fuzzy” query. On the other hand, WebRat’s clustering approach, works best when users do not search for well-defined environmental information objects, but want to get an overview over a topic based on comparable few keywords entered in a textual form. In such cases, WebRat provides an excellent overview and the opportunity to explore a result set by interactively refining the initial query. However, locating a clearly specified environmental object is not the strength of WebRat: A classical query and result list system is superior in such a case because it allows for faster and more accurate identification of single items.

3.4. Visualising cluster hierarchy

As a recent trend, a number of web search engines has started to cluster search results by topic and to display found clusters in tree structures, which allows grouping and viewing results on various levels. While such measures surely increase usability, there are several user-interface optimisations which could be applied, including reduction or elimination of the need for scrolling (as necessary when large branches are expanded in a standard tree view) and displaying a visualisation that is stable (navigating the hierarchy down some levels of depth will not cause any changes in the presentation of the upper levels, as in the case of hyperbolic trees [CITE: HYPERBOLIC]).

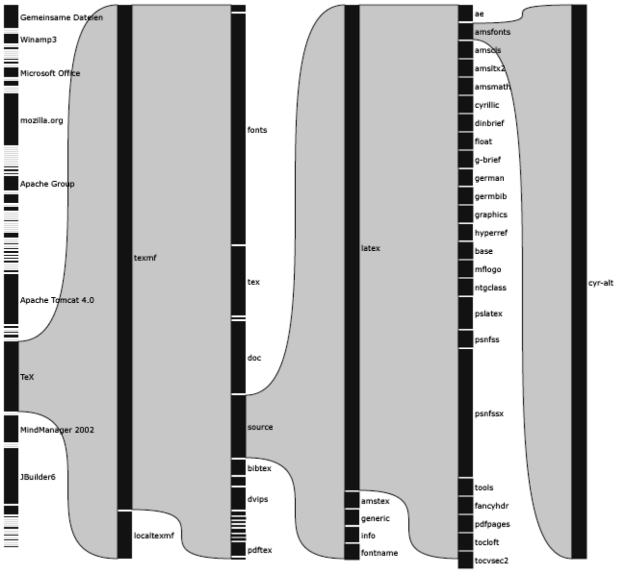

Figure 4: Visualising a file system hierarchy with statistical and topical information

A team at the KnowCenter has developed a clustering and visualisation solution based on the WebRat framework, which meets the aforementioned criteria and enhances standard tree views by incorporating statistical information like relevance, relative position in the tree and amount of data found in a specific node and its sub nodes. The result is an interactive statistical graphic, which is rendered once using the SVG graphics standard [CITE:SVG] and can repeatedly be used to navigate a topical structure.

4. Related Work

4.1. Search Result Clustering

The clustering of search results is a central feature of WebRat. A number of systems now provide similar functionality, extracting structure from content or meta-data.

Lighthouse [CITE:LIGHTHOUSE] is a visual meta-search engine. The search query is sent to several search engines (for example Alta Vista or Google) and search results are retrieved, analysed and thematically organised. A 2D or 3D layout is computed, with spatial proximity indicating similarity inbetween objects. Thematic organisation of documents is not incremental but performed after all required search results have been gathered. The system can handle at most a few hundred documents and the layouts can not be extended incrementally. The use of a very simple, non-expressive visualisation metaphor (spheres in space), the absence of abstract descriptions for groups of related documents as well as the lack of distinct, recognisable group visualisation elements are serious disadvantages of the system.

Vivisimo [CITE:VIVISIMO] allows documents to be assigned to more than one class within the hierarchy to express the complex relations in the retrieved document set, but this might be a potential source of confusion to a non-expert user. The visualisation provided is a classical tree view, with determined forming the nodes of the tree. We believe that a less abstract metaphore would aid users in comprehending and navigating the topcial structure.

WebMap's InternetMap [CITE:WEBMAP] displays search results in the context of a pregenerated hierarchy of web sites. Classes of similar documents are represented by hierarchically nested, bordered regions, where the areas of the regions are a measure for the number of contained documents and sub-classes. Spatial proximity is a measure of relatedness between classes and between documents within a class. The main limitation of the system is that it is static in the sense that the user can only browse and search within the document set which was pre-processed to create the visualisation.

4.2. Thematic Landscapes

A thematic landscape is a visual representation of a large number of documents which are placed in a 2D or 2.1D space in such a way that their spatial proximity in the layout roughly corresponds to their thematic similarity. Thematic landscapes are mostly used to visualise relations in large, unstructured document collections. WebRats standard visualisation, as well as several derived visualisation modes (but not the topic cascades, compare chapter 3), are based on the idea of thematic landscape. Several other retrieval applications share this approach.

The Bead system [CITE:BEAD] is an interactive, multi-participant visualisation system where a 2.1D information-terrain is represented by a triangulating the layout. Bead can handle several thousands of objects, but its visualisation metaphor is minimalistic and not very intuitive. The system allows for a completely free navigation and it offers very little orientation and navigation help to the user.

The SPIRE project [CITE:SPIRE], applies clustering methods such as kmeans to improve the performance of a special multidimensional scaling method (so called Anchored Least Stress) for computing the 2D layout. The system can produce two similar visualisations: Galaxies visualise documents as stars so that groups of related documents appear as galaxies; ThemeView (previously called Themescape) is a further development of the Galaxies visualisation where an undulated terrain is generated by representing documents topical characteristics as sedimentary layers placing them atop of each other and adding them together. The system was successfully applied to document sets with sizes of several hundred thousands.

VxInsight [CITE:VXINSIGHT] is a visualisation system quite similar to SPIREs ThemeView. The layout is produced by a force-directed placement method using a density-grid technique to reduce execution time complexity. The system also offers a Laplacian eigenvectors method which produces very tight clusters so that a subsequent refinement step with the force-directed placement method is still required. VxInsight can handle document sets containing a million or more objects.

Finally, Kohonens Self Organising Maps (SOM) [CITE:SOMS] is an approach based on artificial neural networks. Document term vectors are used to train the neural network which produces a topically organised map by placing the documents into automatically identified 2D subject areas. The sizes of these subject areas as well as their locations and relative positions are determined by thematic relationships of their contained documents. Although SOMs can be applied to compute maps for millions of documents, their main disadvantage is that neural networks have a very slow learning phase.

5. Concluding Remarks

In this paper, we have presented WebRat, a retrieval, clustering and visualisation framework well suited to deal with the large, unstructured knowledge spaces provided by todays web search engines. Several applications of WebRat to a variety of domains have demonstrated that an agile retrieval and visualisation system that quickly adapts to changing user requirements, data sources and availability of meta-information is possible. We hope that WebRat and the applications built upon it will aide knowledge workers in their most crucial tasks: Locating and analysing information with context, to generate new knowledge.

6. Future Work

WebRat is constantly being extended with the addition of further sources and repositories, the query results of which can be merged to extend the number of relevant documents found. Among others, we are exploring the opportunities which organizing eMail or newsgroup archives with WebRat could open up.

Metadata will be integrated more tightly by weighting known metadata fields in comparison to general information provided in content snippets. Another project at the Know-Center dealt with meta-data based retrieval during 2002, and integrating this approach will be a major challenge for the time to come.

7. References

[1] A.B. Smith, C.D. Jones, and E.F. Roberts, “Article Title”, *Journal*, Publisher, Location, Date, pp. 1-10.

[2] Jones, C.D., A.B. Smith, and E.F. Roberts, *Book Title*, Publisher, Location, Date.