Topic Cascades: An Interactive Interface for Exploration of Clustered Web Search Results Based on the SVG Standard

M. Lux, M. Granitzer, V. Sabol, W. Kienreich and J. Becker

Know-Center, Inffeldgasse 16c, 8010 Graz, Austria
{mlux,mgrani,vsabol,wkien}@know-center.at
http://www.know-center.at
jutta.becker@acm.org

Abstract. The WebRat is a light-weight, web-based retrieval, clustering and visualisation framework which can be used to quickly design and implement search solutions for a wide area of application domains. We have employed this framework to create a web meta search engine combined with an interactive visualisation and navigation toolkit. Based on the SVG graphics standard, this application allows users to explore search results in a quick and efficient way, by choosing from topically organized result groups. The visualisation and the cluster representation can be stored and reused. We have combined hierarchical navigation of search result sets with a topical similarity based arrangement of these results in one consistent, standard-based system which demonstrates the potential of SVG for web-based visualisation solutions.

1 Overview

This paper introduces an innovative SVG-based visualisation component of the WebRat retrieval and visualisation framework, for exploration and browsing of clustered web search results. An overview of WebRat and motivation for the chosen visualisation approach are provided. Preliminary results of usability studies and an outlook on future activities are discussed.

2 Introducing WebRat

When searching for a specific topic in the Internet very large amounts of information are returned, and a significant portion of hits is often not at all of interest. No explicit relations between the retrieved documents are returned which makes it difficult to obtain an overview, navigate the search results and find relevant information. WebRat [1] is web-based framework upon which incremental retrieval, clustering and visualisation applications can be built. It addresses the problem of information

V. Palade, R.J. Howlett, and L.C. Jain (Eds.): KES 2003, LNAI 2773, pp. 967-974, 2003.

overload by integrating documents from different data sources into a topically organized presentation and providing means for interactive exploration of the document set. The system has already been applied to web query refinement and metadata based environmental information search [2]. A meta search engine based on the WebRat framework is available at http://www.know-center.at/webrat.

2.1 Retrieval and Processing

WebRat employs an innovative, incremental, three-stage processing of retrieved documents introducing a feedback-loop to improve processing speed and the quality of results. Retrieval begins by sending the search query to various web data sources. In the high-dimensional stage documents retrieved from different data sources are transformed into a language independent term vector representation by using n-gram decomposition [3], and a term-frequency inverse-document-frequency (TFIDF) weighting scheme is applied. Words and double words which were sources for different n-grams build word vectors which are used for key-term extraction. The mapping stage maps the high-dimensional vector representation to the 2D viewport space by employing an incremental force-directed placement algorithm enhanced by a stochastic sampling schema [6] and a clustering method. The computed 2D configuration reflects the high-dimensional relations of the search results (as far as it is possible) - topically similar documents form dense groups in the 2D layout. Low-dimensional stage makes use of advanced rendering techniques to quickly compute a density matrix based on the 2D document coordinates. The density matrix serves two purposes. First, the landscape background image is generated from it where islands represent dense areas of topically similar documents. Second, it is used to periodically identify the density maxima positions in 2D which are used as cluster seeds. Clusters are created by assigning each document to the nearest seed.

A feedback loop is created by passing clusters back to the first two stages. The high-dimensional component computes high-dimensional centroids of the clusters and employs statistical methods to extract key-terms from the underlying documents and compute cluster descriptors. These descriptors are used by the labelling engine to describe groups of topically similar document the user sees on the map. Mapping stage uses clusters to reduce the computational complexity and increase separation in the layout.

2.2 Visualisation

In its basic form, WebRat provides a visualisation following the concept of thematic landscapes, which is a visual representation in which the spatial proximity between visualised entities is a measure for thematic similarity of the underlying documents. Systems like Bead [5] visualise document sets as galaxies of stars or a landscape. However, a thematic landscape is not very appropriate for displaying hierarchically organised information (i.e. clusters). Beside the standard tree visualisation, we also considered cone trees [14], hyperbolic trees (star trees) [17] and information slices [13]. None of these approaches fulfilled our requirements regarding use of screen real estate, ease of use, providing and overview, avoiding occlusion and the necessity to scroll.

Fig. 1. WebRat's landscape visualisation. Query was "SVG"

There is a general trend towards clustering of search results as performed by Alta Vista and Northern Light search engines. The Vivisimo [11] meta search engine creates a tree-like structure displaying a hierarchy of clustered search results. Lighthouse [4] also performs clustering and computes 2D or 3D thematic layouts. Systems like InfoSky [7] try to combine hierarchical structure with visualising document similarity for very large document corpora. Mondeca [10] is doing research on SVG based visualisation of Topic Maps. All considered systems either provide only simple tree representations or present results in a complex 2D or 3D hierarchical layout. They also feature fixed, non-configurable visualisation stages. WebRat provides more flexibility offering a pluggable visualization interface and is more advanced in the sense that its visualisation works incrementally, meaning that it can incorporate results as they arrive.

2.3 Intelligent Retrieval with WebRat

WebRat supports several powerful features in the area of intelligent retrieval and visualisation of search results:

- The capability to retrieve textual documents from a number of heterogeneous data sources, such as search engines, knowledge management environments [16], environmental databases [15], and others. WebRat combines these into a single consistent internal representation.
- Thematical organization of retrieved documents through unsupervised incremental clustering. WebRat accommodates new results as they arrive into a growing, topically organized 2D map, identifies clusters of similar documents and inserts new documents into the identified topical hierarchy.
- Dynamical labeling provides enhanced orientation while navigating the document set. WebRat computes labels on the fly, depending on the zoom level

and hierarchy depth to best describe documents and documents groups the user is currently focusing on.
- Query refinement through recommending additional query terms. Depending on the area the user is focusing on WebRat automatically proposes query terms to narrow down the search query.

Based on these features, a scenario can be developed in which a user issues a specific query, such as “SVG”, gets an overview of the result set and learns about the sub-topics and vocabulary describing each sub-topic. From the user's focus on one special field of interest, for example “ANIMATIONS IMPLEMENTING” (see fig. 1), WebRat “recognizes” the point of interest and automatically extends the search query to gather more documents on that particular sub-topic. As topics overlap between clusters, results of the refined query are incorporated not only within the “implementing animations with SVG” topic, but also within other topics related to SVG (the original query). In conjunction with WebRat's demonstrated ability to present users with a visual summarisation of deep hierarchies [16], this scenario can be extended to the case of a knowledge management system enhancement which, by determining points of interest as described above, offers users a personalised access point of a knowledge space which is created autonomously.

3 The SVG Application

3.1 Motivation

While the features and scenarios described promise several advantages for users querying large repositories, in a real-life environment the 2D visualisation used by WebRat is often not suitable. Most knowledge management systems provide users with a tree view navigation system covering a sizeable amount of screen real estate, with the remaining space holding content and metadata display elements which cannot be hidden or discarded without loosing most of the intended functionality. In consequence, a major challenge has been to develop an interactive visualisation suitable for presenting hierarchies which at the same time offers as much of WebRat's topic-based navigational capabilities and eliminates several of the drawbacks of the standard tree approaches. For example, reduction or elimination of the need for scrolling as experienced with standard tree views when large branches are expanded has been targeted. We also wanted a visualisation that is stable in the sense that navigating the hierarchy down to 4 or 5 levels of depth will not cause any changes in the presentation of the upper levels, as it is the case with hyperbolic trees. The visualisation should offer an overview for several levels of depth, without occlusion and the necessity to scroll, through better use of screen real estate. We also wanted to incorporate statistical information like relevance and size as it is the case in the WebTOC system [18]. However the standard tree is not, to our opinion, a structure where extra information is incorporated and embedded in a clear manner.

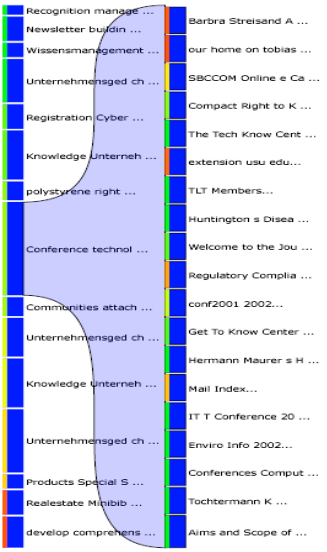

Fig. 2. SVG Visualisation

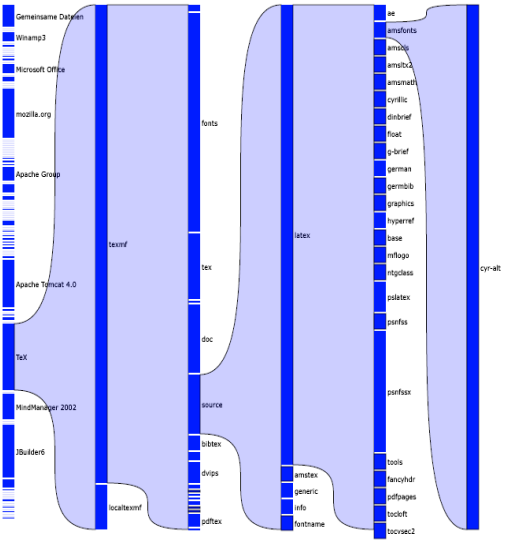

Fig. 3. File system visualization

Our requirements for the format included the following: vector based visualisation, mouse interactivity for tasks like opening an URL, showing a description or showing and hiding parts of the visualisation, and, finally, saving or exporting the results in a reusable and interchangeable format. Searching for standards for presenting graphical information on the Web we identified two possible candidates: Macromedia Flash [8] and Scalable Vector Graphics (SVG) [9]. Although software providers deliver browsers supporting Flash whereas for viewing SVG an extra plug-in is needed, we preferred the second option, because SVG is new, state-of-the-art, XML-based and, compared to Flash, an open standard.

3.2 Visualisation Metaphor

For visualisation a modified tree representation is used. On the one hand the obvious advantages of a tree for navigating and grouping the results could be retained, on the other hand the modifications allowed us to integrate statistical information about the results or group of results itself. The nodes of the tree are represented by blue rectangles, where inner nodes represent clusters and the leafs represent the actual search results (see fig. 2). Additional data is integrated in the visualisation in three different ways. Firstly, the coloured bars attached on the left side of each rectangle indicate the relevance of the results. Relevance is coded by colours, from green across yellow to red, where green represents high and red low relevance. Secondly, the height of the rectangles: it indicates the cardinality of the cluster. The third kind of information embedded in the visualisation are the relative positions of the rectangles. Neighbouring rectangles have a higher content based similarity than results which are not adjacent.

3.3 Application

As an application for our SVG visualisation, we have implemented a web meta-search engine based on the WebRat framework and added the SVG visualisation as a result browser. Users access the system through a standard web search interface featuring an input box for query terms and some additional controls for configuration of a query session. After starting a query, the WebRat system queries the search engines, retrieves, transforms and organises the results and then starts the SVG component in a new browser window, sending to it the results of the query in XML format. Finally, the uppermost level of clusters is displayed in the new browser window and the user can start exploring the search results. Moving the mouse cursor over a blue rectangle results in a color change to red. Clicking on the rectangle representing a group opens a new column, displaying its children. The leafs of the tree are representing single search results and clicking on one of them will open the URI of that specific search result. Our SVG application creates a persistent visual description of a search result space which can be saved and recalled for later use.

4 Future Work

In its current form, the visualisation degrades in usability if too many levels are open, because the screen gets cluttered with too much information. We plan to address this issue by fading unused branches or eventually introducing pan and zoom operations. The descriptions of the leafs and nodes are simply cut off after a specific number of characters. Even though the full description of the node is displayed in a status line after clicking on it, a mouse over effect should reveal the full description.

5 Conclusion

While search result visualisations such as information landscapes are an improvement over the standard ranked list, a more structured and discrete approach can be of benefit in certain situations. We have combined classical tree-based navigation with representation of topical similarity to create a visual representation of search result sets which is both easy to navigate and, at the same time, expresses additional properties of the result set such as topical similarity and statistical information. We found SVG to be a stable and elegant standard for vector graphics which could prove quite useful in developing web-based visualisation applications.

Acknowledgements

The Know-Center is a competence center funded within the Austrian Competence Center Programme K plus under the auspices of the Austrian Ministry of Transport, Innovation and Technology (www.kplus.at).

References

[1] Sabol, V., Kienreich, W., Granitzer, M., Becker, J., Tochtermann, K., Andrews, K. (2002). "Applications of a lightweight, web-based retrieval, clustering and visualisation framework", in "Proceedings of the 4th International Conference on Practical Aspects of Knowledge Management", Vienna University of Technology, Austria.

[2] Tochtermann, K., Sabol, V., Kienreich, W., Granitzer, M. and Becker, J. (2002). "Intelligent Maps and Information Landscapes: Two new Approaches to support Search and Retrieval of Environmental Information Objects", in Proceedings of 16th International Conference on Informatics for Environmental Protection, Vienna University of Technology, Austria.

[3] Cavnar, W.B., Trenkle, J. M. (1994). "n-Gram based text categorization". In Symposion on Document Analysis and Information Retrieval, p161-176, University of Nevada, Las Vegas.

[4] Leuski, A., Allan, J. (2000). "Lighthouse: Showing the Way to Relevant Information." In Proceedings of IEEE Symposium on Information Visualisation 2000, pp. 125-130, InfoVis2000, Salt Lake City, Utah.

[5] Chalmers M. (1993). "Using a landscape metaphor to represent a corpus of documents." in Proceedings European Conference on Spatial Information Theory, COSIT 93, pages 337-390, Elba.

[6] Chalmers M. (1996). "A linear iteration time layout algorithm for visualising high-dimensional data." in Proceedings Visualization96, IEEE Computer Society, pages 127-132, San Francisco, California.

[7] Andrews, K., Kienreich, W., Sabol, V., Becker, J., Droschl, G., Kappe, F., Granitzer, M., Auer, P., Tochtermann, K. (2002). "The InfoSky Visual Explorer: Exploiting Hierarchical Structure and Document Similarities." In Palgrave Journal on Information Visualisation, Hampshire, England.

[8] Macromedia Inc. (2003). "Macromedia Flash Support Center", http://www.macromedia.com/support/flash/

[9] W3C, SVG Working Group (2003) "Scalable Vector Graphics", http://www.w3.org/Graphics/SVG

[10] Delahousse, J.: "Index and knowledge drawing: a natural bridge from Topic Maps to XML SVG", Mondeca, 2001, http://www.idealliance.org/papers/xml2001/papers/html/04-04-02.html

[11] Vivisimo Inc. (2003) "Vivissimo Document Clustering", http://www.vivissimo.com

[12] Hunter, J. (2003). „JDom", http://www.jdom.org

[13] Andrews, K., Heidegger, H. (1998) "Information Slices: Visualising and Exploring Large Hierarchies using Cascading, Semi-Circular Discs" In Late Breaking Hot Topic Paper, IEEE InfoVis'98, Research Triangle Park, North Carolina.

[14] Robertson, G. G., Mackinlay, J.D., Card, S.K. (1991) "Cone trees: Animated 3D Visualisations of Hierarchical Information." In Proceedings CHI91, pages 189-194, New Orleans, Louisiana.

[15] Tochtermann, K., Sabol, V., Kienreich, W., Granitzer, M., Becker, J. (2003). ”Enhancing Environmental Search Engines with Information Landscapes", ISESS - 8th International Symposium on Environmental Software Systems, Vienna, Austria

[16] Kienreich, W., Sabol, V., Granitzer M., Becker J., Tochtermann K. (2003). "Themenkarten als Ergänzung zu hierarchiebasierter Navigation und Suche in Wissensmanagementsystemen", 4. Oldenburger Forum Wissensmanagement, Oldenburg, Germany.

[17] Lamping, J., Rao, R. (1994) “Laying out and visualizing large trees using a hyperbolic space” In Proceedings UIST94, p. 13-14, Marina del Rey, California.

[18] Nation, D. A., Plaisant, C., Marchionini, G., Komlodi, A. (1997) “Visualizing websites using a hierarchical table of contents browser: WebTOC”, University of Maryland.