New Paradigms in Information Visualization

Peter Au, Matthew Carey, Shalini Sewraz, Yike Guo, and Stefan M Rüger
Dept of Computing, Imperial College, London SW7 2BZ, England (http://www.doc.ic.ac.uk/~srueger)

Abstract. *We present three new visualization front-ends that aid navigation through the set of documents returned by a search engine (hit documents). We cluster the hit documents to visually group these documents and label the groups with related words. The different front-ends cater for different user needs, but all can browse cluster information as well as drilling up or down in one or more clusters and refining the search using one or more of the suggested related keywords.*

Introduction

Broad one or two-word searches in conventional search engines are often plagued by low precision returning thousands of hit documents as their output. A common problem with this is that users have to wade through much non-relevant material before finding relevant documents. A large set of search results could be narrowed by *query refinement* which means augmenting the query with additional search terms after the initial search. This is in fact another interesting and well-known recent advance of AltaVista, called Live Topics. This feature is useful in narrowing down a search; however, the search result is still shown as a long list of pages to browse through. In our opinion, the strategy should be to shift the user's mental load from these slower thought-intensive processes such as reading to faster perceptual processes such as pattern recognition in a visual display. The ranked-list metaphor, though simple, is too restrictive: with large volumes of data displayed on multiple pages we find ourselves searching all over again! Furthermore, in conventional search engines the documents are ultimately ranked with the aim to order them according to relevance to the user. This appears to be overly ambitious as even advanced ranking algorithms cannot know whether the user prefers documents about "hardware" or "software" when the query simply was "computer".

We suggest clustering the hit documents and make use of the obtained groups with interactive displays. We have overcome the curse of dimensionality by representing each hit document with a small vector that is a histogram of related terms such as "software", "UNIX", "IBM", "users" for the query "computer". We compute these related terms dynamically at query time from the subset of hit documents, ie, when the query is refined to "computer hardware", a new set of related words emerges; for details see [9]. The obtained clusters based on this representation have proven to be meaningful in experiments with a large database of TREC human relevance judgements [12].

The recent decade has seen much interest in information visualization, see eg [7, 4, 11, 5, 1, 2]. In the following we suggest three new paradigms in information visualization displaying clusters of documents.

1 Sammon Cluster View

In this paradigm, we use Sammon mapping [8] to convert the high-dimensional cluster centroid vectors to two dimensions, while trying to preserve the distance among the clusters. These 2-dim cluster vectors will ultimately be mapped onto the interface, thereby providing a visual landscape for navigation. Clustering cannot be performed in advance on the collection as a whole, as the features that encode a document are the related words which depend on the query (indeed, clustering should not be performed in advance, as the hit documents returned by a query should ultimately determine how these documents are best projected). It is to be noted that bringing higher dimensionality down to lower dimensionality for displaying is a trade-off between precision and cost. Lower dimensionality means somewhat rougher representations of document relationships but cheaper access and manipulation, the latter of which is more important here.

Each cluster is represented by a circle on the screen, see Figure 1. The size and color shade indicate the size of the cluster. Color is not used in the moment and left for custom implementations. The distance between any two circles in the panel represents the similarity of their respective clusters: the nearer the clusters, the more likely the documents contained therein will be of similar context thereby enabling the user to rapidly find all similar documents. It is not uncommon for a session to move across the spectrum from browsing to searching. Indeed, each new piece of information seems to give new ideas and directions to follow, and consequently, a conceptualization of the query [3]. For either purpose, it was deemed useful to provide a *tooltip box* which contains additional information about each cluster (such as the top-five related words of this particular cluster and the number of documents) and which appears when the mouse cursor is positioned over a cluster. Also, operations such as *select, drill up* and *drill down* can be executed for this particular cluster or selection of clusters.

Keyword refinements are possible within clusters or across a selection of clusters. When browsing through the clusters and identifying an interesting cluster, the user will

SIGIR 2000 7/00 Athens, Greece
© 2000 ACM 1-58113-226-3/00/0007...$5.00

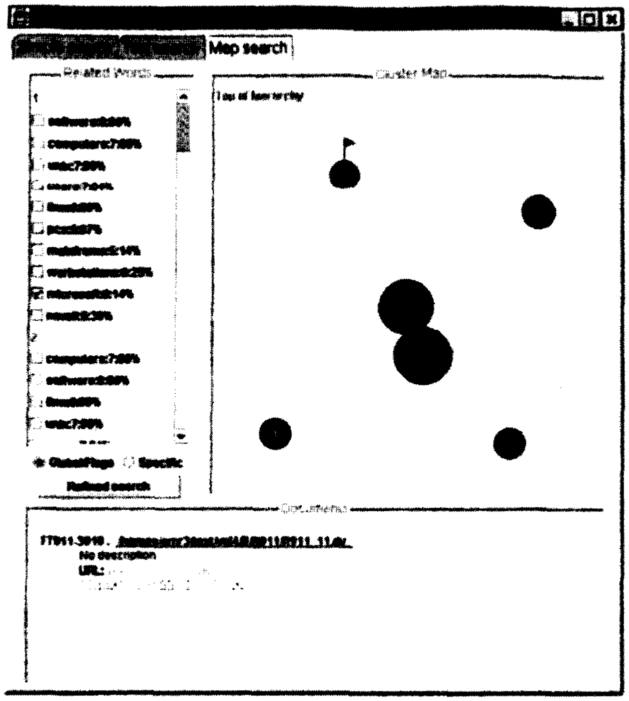
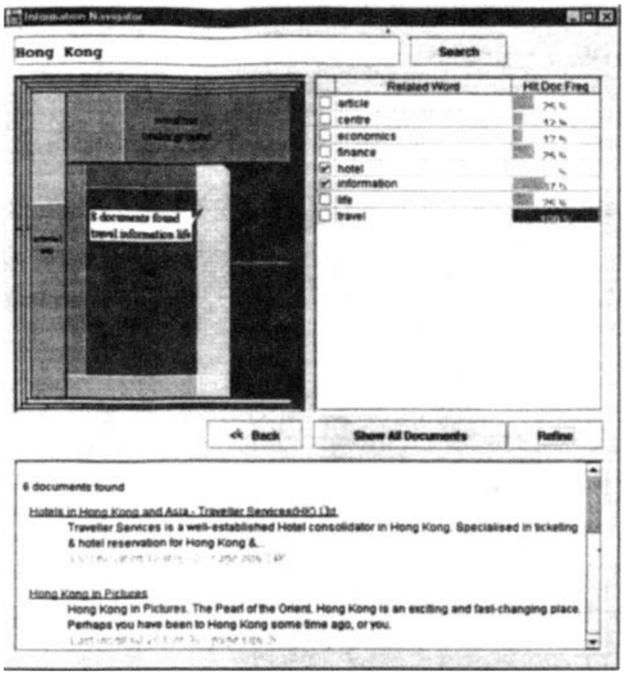


Figure 1: Sammon cluster view



	Related Word	Hit Doc Freq
□	article	[illegible]
□	centre	[illegible]
□	economics	[illegible]
□	finance	[illegible]
☑	hotel	[illegible]
☑	information	[illegible]
□	life	[illegible]
□	travel	[illegible]

6 documents found

Hotels in Hong Kong and Asia - Traveller Services(HK) Ltd

Traveller Services is a well-established Hotel consolidator in Hong Kong. Specialised in ticketing & hotel reservation for Hong Kong &...

Hong Kong in Pictures

Hong Kong in Pictures. The Pearl of the Orient. Hong Kong is an exciting and fast-changing place. Perhaps you have been to Hong Kong some time ago, or you...

Figure 2: Tree-map focus+context approach

probably want to see the document titles and descriptions contained in that cluster. As this should be a quick action, simply clicking on a cluster will display in the bottom panel a list of the document titles, descriptions and URLs. Similarly, a user might want to see at the same time the list of document titles and descriptions for more than one cluster. As this will be done after some thought and selection of relevant clusters, this option is in the tooltip box, and upon clicking on *display selection* this information will be displayed in the bottom panel.

2 Tree-map focus+context approach

The navigator mainly consists of three panes: a clustering view, a related-words table and a document display. When a user types a string to search, the navigator will visualize the clustering result into different rectangles based on an idea found in [10]. Each rectangle contains a set of documents, see Figure 2. Similar clusters are grouped into a category, which is represented by a bounded box. Each bounded box has some labels, normally the top ranked related-words among the underlying clusters. The two levels of clusters are brought about by cutting the initial dendrogram of our buckshot clustering at two similarity levels.

A tool-tip box is displayed when a user puts the mouse over a candidate cluster. It briefly describes the cluster, such as the hit-document rate and some top ranked related-words. Moreover, users can inspect all the related words within the cluster by clicking the left mouse-button. A table contains those related-words and their hit-document frequency will be displayed on the right-hand side. A small triangle is marked on the upper right corner of the rectangle so as to users can easily identify which cluster the table is referring to. Users can check the relevant related-words in the table and the navigator will retrieve documents that are similar to those pointed out by them. The novel feature of the navigator is that users can drill down on a selected cluster by clicking the right mouse-button. All document references within the cluster will be sent back to the back-end to redo the clustering; the current rectangles will scatter to the edge and a new clustering map will then display. Each frame implies the clustering map has been drilled down in one level. Users can restore the previous view by clicking the *back* button.

3 Radviz interactive visualization

This is work based on ideas found in [6], where the related words are initially arranged on a circle and connected with an invisible spring to each document they appear in. The documents are thus placed at an equilibrium between their related words and the centre of the circle, see Figure 3. Hence, we make direct use of the related-word hit-document matrix without explicit clustering.

Moving the mouse over an x-cross representing a document shows a bubble that displays basic information about the document (in our preliminary implementation just the document name); at the same time the related words that appear in the document are highlighted. Moving the mouse over a related word highlights all documents that use this related word. Upon clicking on the

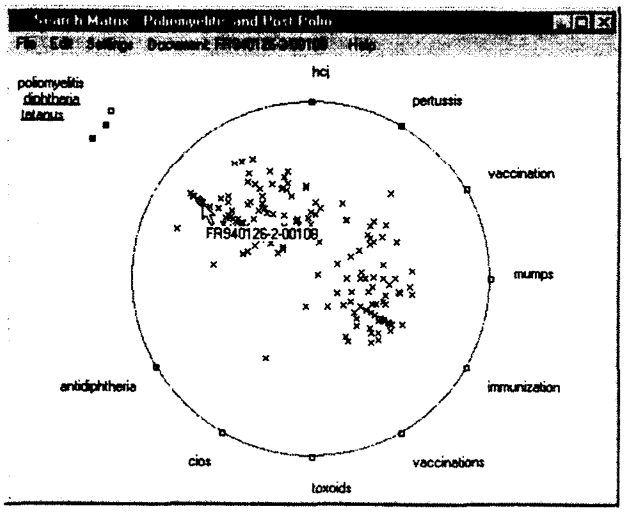


Figure 3: Radviz interactive visualization

related word, it can be dragged freely over the screen and the corresponding highlighted documents follow the movement. In this way, groups of interesting documents can be formed interactively by the user. Pull down menus let the user chose the number of related words to work with or let the user manually generate the subset of related words to look at. Another, automatic way of seeding the related words on the circle is clustering them using the *columns* of the word document matrix (note that document clustering is done usmg the *rows* of this matrix). Dimensionality reduction for these column vectors could be achieved by restricting the representation of a related word to the best-ranked k, say 30, documents returned by the search engine. In this way, "mumps," "measles," and "rubella" are made neighbors on the initial circle and a clustering of the documents on the screen follows more or less naturally from the seeding of the related words.

Conclusions

Our work has contributed to the visualization and browsing of the set of document returned by a search engine in a number of ways. 0) In earlier work, we identified relevant features of this document set, the related words; these are used for dimensionality reduction & improved clustering, cluster labelling, query refinement and visualization. 1) Using Sammon's algorithm we are able to create a setting with a holistic view giving primarily information about a first-order cluster structure and inter-cluster relations. The main purpose is to quickly weed out irrelevant clusters and drill down m one or more relevant clusters. 2) Using the tree-map algorithm, we are able to display second-order cluster structure at one glance. Applications include learning about the fine-structure and nature of queried object as coded in the actual use of the related words in the document repository - ideal for a user knowing not much about the subject. 3) A related words clustering with the Radviz visualization gives rise to another novel document clustering approach, one where the user can control the building of groups by interactively moving the related words around. We feel that this interface is particularly useful for an experimental, user-driven approach to form clusters and to get a suitable ranking by interactively moving the related words around on the screen.

References

[1] M Ankerst, D Keim, and H Kriegel. Circle segments: A technique for visually exploring large multidimensional data sets. In *IEEE Visualization '96*, 1996.

[2] J Assa, D Cohen-Or, and T Milo. Displaying data in multidimensional relevance space with 2d visualization maps. In *IEEE Visualizatıon '97*, 1997.

[3] M Baldonado and T Winograd. Sensemaker: An information-exploration interface supporting the contextual evolution of a user's interests. *CHI Electronıc Proc*, 1997.

[4] M Chalmers and P Chitson. Bead: Explorations in information visualisation. In *Proc of the 15th Intl ACM SIGIR Conf*, 1992.

[5] M Hemmje, C Kunkel, and A Willet. Lyberworld - a visualization user interface supporting fulltext retrieval. In *Proc of the 17th Intl ACM SIGIR Conf*, 1994.

[6] P Hoffman and G Grinstein. Dimensional anchors: A graphic primitive for multidimensional multivariate information visualizations. In *Proc of the NPIV 99*, 1999.

[7] R Korfhage. To see or not to see - is that the query? In *Proc of the 14th Intl ACM SIGIR Conf*, 1991.

[8] J W Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18(5), 1969.

[9] S Sewraz and S M Rüger. A visual information-retrieval navigator. In *Proc of the BCS IRSG2000*, 2000.

[10] B Shneiderman. Tree visualization with Tree-Maps: 2-d space-filling approach. *ACM Transactıons on Graphıcs*, 11(1):92–99, 1992.

[11] A Spoerry. Infocrystal: A visual tool for information retrieval & management. In *Proc of Informatıon, Knowledge and Management 93*, 1993.

[12] G Zervas and S M Rüger. The curse of dimensionality and document clustering. In *Proc of the IEE Searching for Information: AI and IR Approaches*, 1999.

Acknowledgements: This work was partially supported by the EPSRC, UK, and Fujitsu Laboratories of Japan. The authors thank Christian Brenninkmeijer, Athina Economou and Philip Street for their contribution to the initial design study of the Radviz visualization.