

## ClusterBook, a Tool for Dual Information Access

Gheorghe Mureşan, David J. Harper, Ayşe Göker and Peter Lowit  
School of Computer and Mathematical Sciences  
The Robert Gordon University  
Aberdeen AB25 1HG, UK  
{gm,djh,asga,pl}@scms.rgu.ac.uk

ClusterBook is an interface that allows dual access to a document collection by combining a hierarchic, structural view, based on clustering the collection, with a linear view, based on ranking the collection relative to a query.

The user can explore the collection by using the structure to:

- browse the clusters of interest, looking for particular documents, or explore the information space
- become familiar with the collection, with its concepts and vocabulary
- disambiguate terms and concepts, based on the context
- assess the relevance of documents, based on the context

The retrieval strategies supported by ClusterBook are:

- **best-match search** - the system returns in the *ranked viewpanel* a list of documents ranked according to the expected relevance to the user's query
- **browsing the structured collection** - in the *overviewpanel*. The starting point can be a cluster or a document identified by a search, or simply the root of the hierarchy (tree). The user can expand branches of interest, looking for more detail, or can collapse branches that look un-interesting, in order to reduce the information load and better use the display.
- **cluster-based searching** - the system returns in the *ranked viewpanel* a list of clusters that best match the user's query and also highlights them in the *overviewpanel*. From a cluster selected as a starting point, the user can choose to perform a top-down, bottom-up or horizontal search. The default search is top-down, from the top of the hierarchy.

The user cannot only choose among these strategies, but also combine them. For example, interesting documents found by a best-match search can be starting points for bottom-up cluster based searches. Also, a cluster identified by a cluster-based search can have its documents ranked by a best-match search. At any point, the user can explore in detail a cluster or document of interest in the *local viewpanel*.

The interface supports the user in integrating the two views of a document collection, hierarchical and linear, by highlighting in the overview panel documents selected by the user in the ranked view panel and vice-versa. Supposing the user employs best-match searching in order to identify documents relevant for an information need, the distribution of these documents in the overview will indicate 'hotspots' of relevant documents and will encourage the user to explore these spots by browsing and, possibly, cluster-based searching.

ClusterBook is an alternative front-end to the WebCluster system [2], as the user's exploration of a collection and the formulation and refinement of an information need can be a first step in a mediated access session to a heterogeneous collection [1].

ClusterBook can also be used as a research tool. An IR researcher can use it to visually verify the **cluster hypothesis**, by observing the distribution of relevant documents in the cluster structure or the distribution of the documents of a cluster in the ranked list of hits. Work with test collections is supported by colour-coding relevance judgements.

**Acknowledgement.** The work was sponsored by Ubilab, Union Bank of Switzerland, Zurich.

### References:

- [1] D.J. Harper, M. Mechkour and G. Mureşan. Document clustering for mediated information access. *Proceedings of the 21<sup>st</sup> Annual BCS-IRSG Colloquium*, Glasgow, April 1999.
- [2] G. Mureşan, D.J. Harper and M. Mechkour. WebCluster, a tool for mediated information access. *Proceedings of SIGIR '99*, Berkeley, August 1999.