ClusterBook, a Tool for System-Mediated Access via Clustered Collections

Gheorghe Mureșan, David J. Harper and Ayşe Göker
School of Computer and Mathematical Sciences
The Robert Gordon University
Aberdeen, UK
{gm, djh, asga}@scms.rgu.ac.uk
http://www.scms.rgu.ac.uk/research/ir

The work is part of the WebCluster project, sponsored by Ubilab, Zurich.
Abstract

In the WebCluster project, we use document clustering to provide structured content portals to large heterogeneous document collections, such as the WWW. The basic idea is to cluster automatically a (relatively) small homogeneous source collection of documents covering some domain of interest. This structured source collection is subsequently used to "mediate" access to larger heterogeneous target collections via search engines, e.g. Informia. The demonstrated tool, ClusterBook, enables a user to browse and search a structured source collection, and to discover relevant clusters of source documents. Based on these relevant clusters, and on the intended target collection, the system can generate good queries that better capture the user's information need. Thus, we mediate access to the target collection(s) (via a search engine) using queries generated after interaction with a structured source collection(s). Both bench and user experiments have shown the potential efficacy of this approach. You are referred to the companion paper presented at ECDL 2001 for further details.

In the demonstration, we present the library metaphor on which system-mediated access is based, and the resultant search process. The ClusterBook interface is demonstrated, and the underlying software architecture is described briefly. Conclusions are drawn concerning the mediated access concept, and the particular tool that implements the concept, ClusterBook.
The library metaphor

1. Select library
2. Consult catalog
3. Browse shelves
4. Use inter-library scheme

Information Need Formulation

Information need
1. Select source collection

2. Explore source collection with ClusterBook

3. Search WWW

Information need

Results

Information Need Formulation

The model of the search process
ClusterBook Interface

Source Collection

Overview of Clustered Source Collection - Folder/Book metaphor

Target Collection

Ranked View - applied to query over Source Collection (can be used for query over Target)

Local View
The client-server architecture

Client (ClusterBook)

Query:  

1.  
2.  
3.  
4.  
5.  
6.  
7.  

Mediated query:  

Server

Document collection  

Vocabulary, index, inverted file  

Cluster hierarchy  

GetCollections  

GetClustering  

SearchClusters  

SearchDocuments  

GetClusteredHits  

GetHits  

Indexer  

Clustering Framework  

Cluster-based searcher  

Ranked-based searcher  

Informia meta-search engine  

GetHits  

Client (ClusterBook) Server