Demonstration of Hierarchical Document Clustering of Digital Library Retrieval Results

C.R. Palmer
J. Pesenti
R.E Valdes-Perez
M.G. Christel
A.G. Hauptmann
D. Ng
H.D. Wactlar

Informedia Project
Carnegie Mellon University
Pittsburgh, PA 15213

valdes@cs.cmu.edu
(corresponding author)

ABSTRACT
As digital libraries grow in size, querying their contents will become as frustrating as querying the web is now. One remedy is to hierarchically cluster the results that are returned by searching a digital library. We demonstrate the clustering of search results from Carnegie Mellon's Informedia database, a large video library that supports indexing and retrieval with automatically generated descriptors.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *clustering.* H.3.7 [**Information Storage and Retrieval**]:Digital Libraries – *user issues.*

General Terms
Design, experimentation.

Keywords
Hierarchical document clustering.

1. INTRODUCTION
As digital libraries grow, accessing these contents will become unwieldy. Systems will output hundreds or thousands of matches, but users will typically examine a handful within the first couple of result pages. The matches that best satisfy an information need may be buried in later pages that are never viewed. These problems are well known; the question is what to do about them.

2. GENERAL APPROACH
We have developed hierarchical *conceptual clustering* algorithms using ideas from artificial intelligence. The conceptual clustering idea advocates that cluster descriptions should be a primary factor *during the formation of clusters*: good clusters possess good concise descriptions. Thus, a cluster is bad if it does not allow for a good description, even if its internal cohesiveness and external distinctiveness are high according to some distance measure.

The basic clustering software is written in C and accepts an input stream of documents in XML format and outputs an XML representation of the hierarchy. The hierarchical folders are

 JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA. ACM 1-58113-345-6/01/0006.

labelled with either single words (if they are informative) or multi-word phrases.

The C clustering software has a number of parameters, the most used of which is a customizable *stoplist*. For example, the term *beginspeechrecognition* often gets inserted in the Informedia news video transcripts, which are captured in textual form via speech recognition methods. This word is certainly not informative, so it would be placed on the stoplist whose members are words or phrases that can be ignored.

3. CLUSTERING NEWS VIDEO TRANSCRIPTS (INFORMEDIA)
The Informedia Project at Carnegie Mellon University began studying the use of speech, image and natural language processing for improving search and discovery in the video medium in 1994. The project has amassed a large video library covering tens of thousands of stories, with indexing and retrieval supported by automatically generated descriptors [1]. The project first focused on better documentation and representation for stories. The user could view a title, a thumbnail image, or a storyboard set of images, and quickly determine relevance based on these abstractions. As the library grew, the need to summarize sets of stories increased. Most Informedia library queries now return hundreds of stories, too many for linear viewing.

We use a Windows Explorer interaction style for browsing search results, which are presented on a split screen: the left contains the browsable cluster (folder) hierarchy and the right frame contains the subset of search results corresponding to the currently activated folder. The two modes of presenting information complement each other well, according to anecdotal and introspective experience (user studies are underway).

The system (http://montblanc.se.cs.cmu.edu/informedia.html) can be run remotely. Currently, most of the processing delay is due to accessing the database, not to the actual clustering, which is fast.

4. ACKNOWLEDGMENTS
Vivisimo, Inc. has made available some of its intellectual property (code and algorithms) to the authors at Carnegie Mellon.

5. REFERENCES

[1] Wactlar, H., Christel, M., Gong, Y., and Hauptmann, A. Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library. IEEE Computer, 32, (Feb 1999), 66-73.