The MetaCombine Project

http://www.metacombine.org/

Aaron Krowne, Stephen Ingram, Saurabh Pathak, Martin Halbert

{akrowne, singram, spatha2, mhalber}@emory.edu

Emory University General Libraries

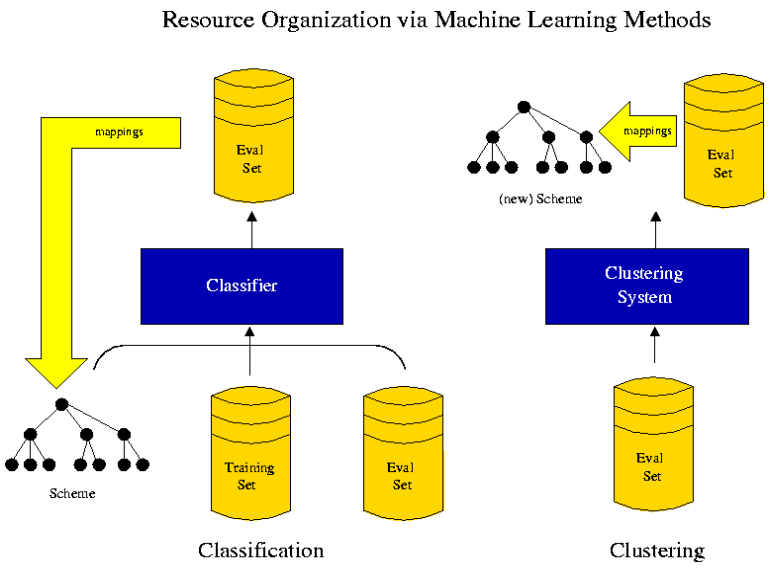
Atlanta, GA

THE METACOMBINE PROJECT

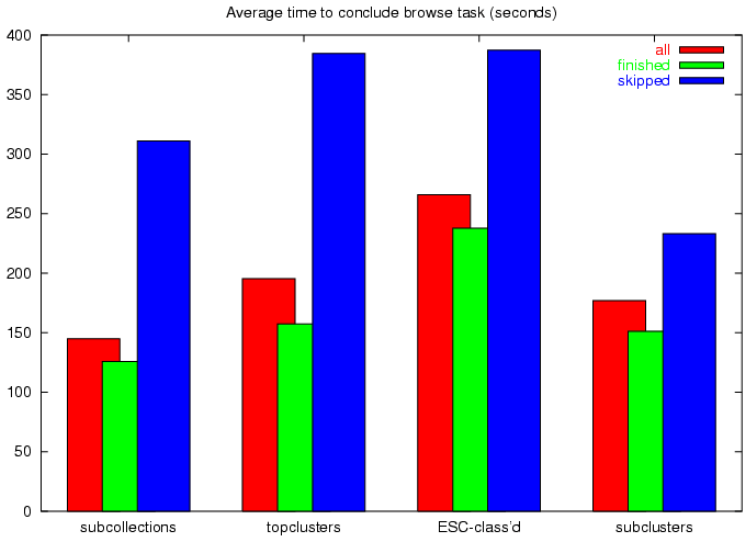
The MetaCombine project is a Mellon-funded effort based at Emory University, with the goal of discovering and developing systems and methods to more meaningfully combine digital libraries, digital library resources, and digital library services. MetaCombine continues within the thread of Emory's MetaScholar digital library initiative. The American South project, the key previous effort of this initiative, demonstrated the use of the (then novel) Open Archives Protocol for Metadata Harvesting (OAI-PMH) to construct a "scholarly portal"-style digital library. This digital library of the history and culture of the American South (at americansouth.org) was built around a union catalog, constructed from the OAI-exposed holdings of collections at other participant institutions. One of the findings of this project was that there is a difficulty in organizing the records collected in such a digital library. Many other similar systems utilizing OAI-harvested collections have encountered the same problem. In the MetaCombine project, we are applying semantic clustering methods to address this problem.

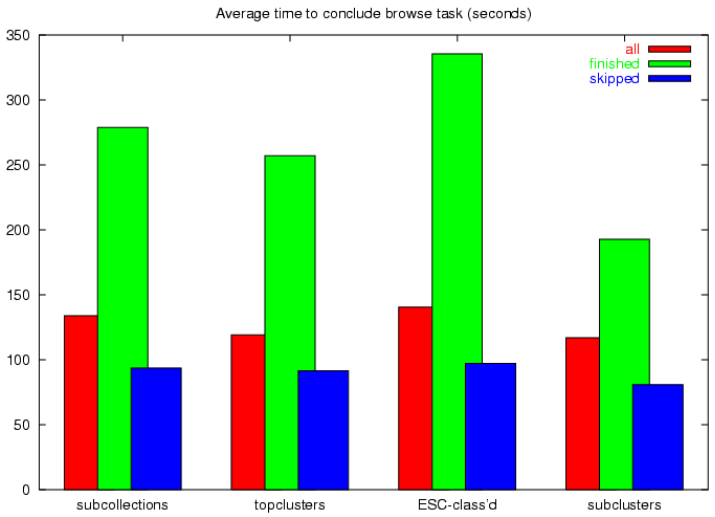
SEMANTIC CLUSTERING

The backbone of the MetaCombine project is semantic clustering, a family of machine learning methods which are used to make connections between records based on the meaning of their content. Within the context of this project, we use the term "semantic clustering" for both text classification and clustering algorithms. The distinction between the two is the presence or absence of an *a priori* ontology (or classification scheme) for the resources. In the case of clustering methods, no classification scheme is pre-provided, and the output of the clustering system is a novel scheme, as well as mappings of the records to the scheme. In our experiments, we tested both the use of classification with a pre-existing ontology as well as clustering to discover a new ontology. For clustering, we used an algorithm based on Non-negative Matrix Factorization (NMF). For classification, we used the BOW system from Carnegie Mellon, with a Rochio online-linear algorithm.

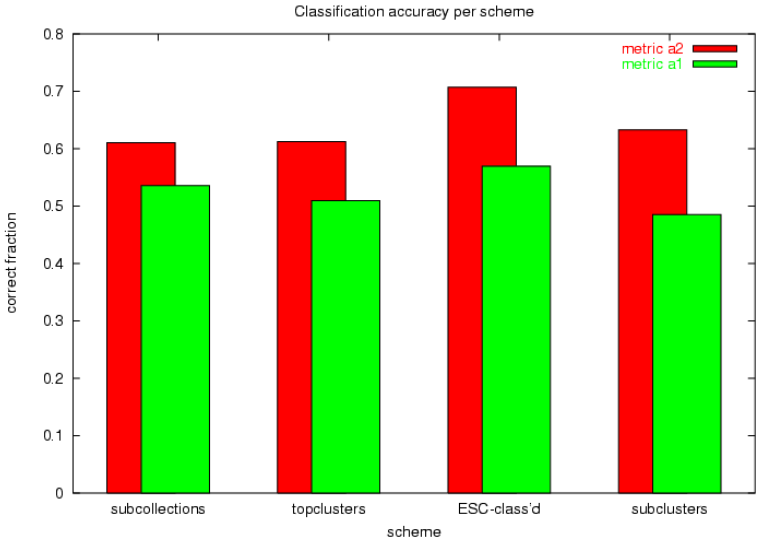


Above: A schematic of the usage of the use of clustering and classification for DL resource organization. Note the varying inputs and outputs. Thus, the presence of a pre-existing organizational scheme and training set, or the need for a new organizational scheme bear heavily on which method may be most applicable or appropriate.

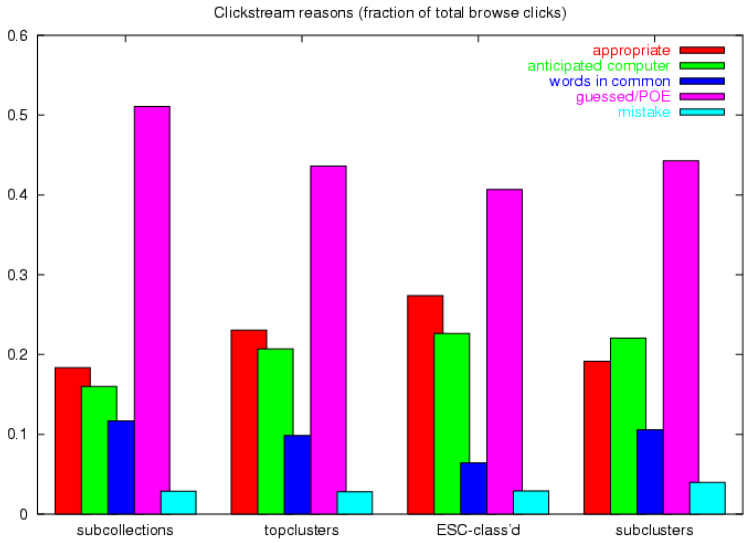




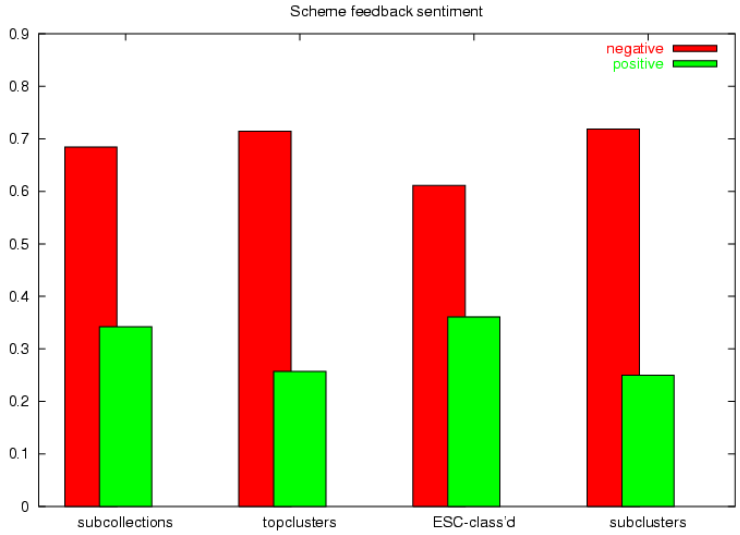
Above two charts: Elapsed time for browse task completion, for all, finished, and skipped tasks. First and second runs of experiment, respectively. Note that all schemes do better than subcollections in at least one of the experimental runs, except ESC-classified.



A summary of classification accuracy of the schemes, which is essentially the fraction of items properly categorized within the scheme. Calculated from navigation feedback given after completion of browse tasks.



An *a posteriori* tally of reasons for navigation choices, given by users after completion of browse tasks. Higher is better for the first three reasons, lower is better for the last two.



A rough tally of negative and positive free-form comments for the browse schemes, given after completion of all browse tasks.

THE EXPERIMENT

We desired to test semantic-clustering derived browse ontologies in a real-world situation. To do this, we used real content (the Americansouth.org records) and real user populations (scholarly experts, college students, and library professionals). There were important characteristics of the Americansouth.org content that could influence the results and which necessitated our tests: imbalanced subcollections, uneven coverage of subject areas, variations in verbosity of metadata, variations in subcollection organizational schemes, and variations in interpretation of metadata fields. All of these factors influence the ability to organize metadata records in practice.

Our experimental method was as follows. Users logged into a web-based experimental system and were presented with resources selected randomly from the Americansouth.org collection. They were instructed to find the resource by browsing through various classification schemes. While they browsed, our system recorded the elapsed time and the navigation path through the scheme. After completing all the browsing, users were asked to give feedback about why they made the navigation decisions they did. The results presented here are from our analysis of the data we captured.

Note: Due to space constraints, we present here only highlights of the results, which come from two separate runs of this experiment.

BROWSE SCHEMES

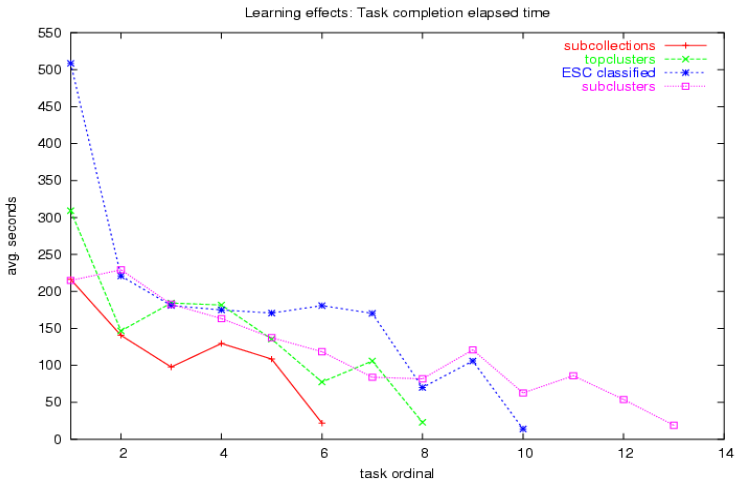
We evaluated four browse schemes in our experiment. One was our control scheme and was conventional in nature. Two were completely AI-based schemes. One was a hybrid of conventional and AI components. The schemes were:

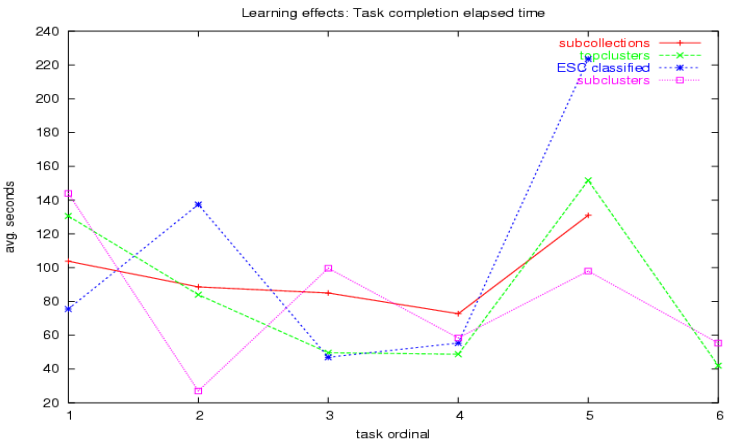
subcollections – The control (or "trivial") scheme. This refers to a simple subdivision of resources based on the collection they were harvested from, and is the easiest way to organize heterogeneous resources in a union catalog. In the results, please observe performance of the schemes *relative* to this one.

topclusters - "Top-level" clusters, a one-level scheme of 25 categories derived from NMF clustering run over the Americansouth.org content.

ESC-classified – An ontology derived from the Encyclopedia of Southern Culture (ESC) subject areas, coupled with classification of resources into this ontology. As a training set, we used the full text of the ESC articles (about 1,400 of them).

subclusters – Subcollection clusters, a hybrid of subcollections and topclusters. The top level of this scheme was identical to subcollections, but within each of these categories was a second level derived from an NMF clustering run for just that subcollection's content.



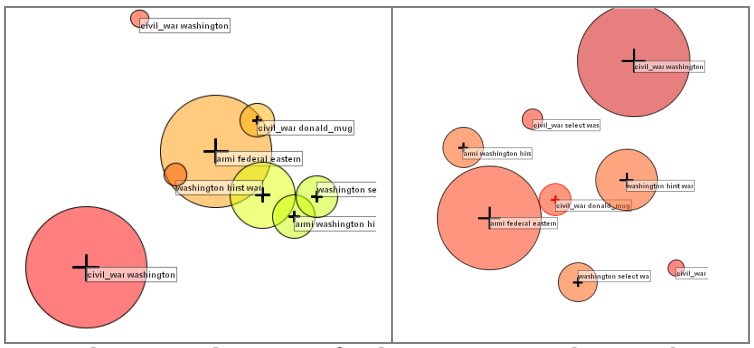


Above two charts: time-to-complete browse tasks, by task order (for first and second run of experiment respectively). This enables us to see the progress of users in learning each scheme. Such a learning effect is visible for all schemes, to varying extents.

CLUSTER VISUALIZATION EXPERIMENTS

We have experimented with the visualization of AI-derived cluster data. Individual clusters' feature vectors are plotted by reducing their dimensionality using dimensionality reduction schemes such as Principal Components Analysis (PCA) or Nonnegative Matrix Factorization. Aside from their geometric position, clusters are labeled for identification and sized according to the number of documents they describe.

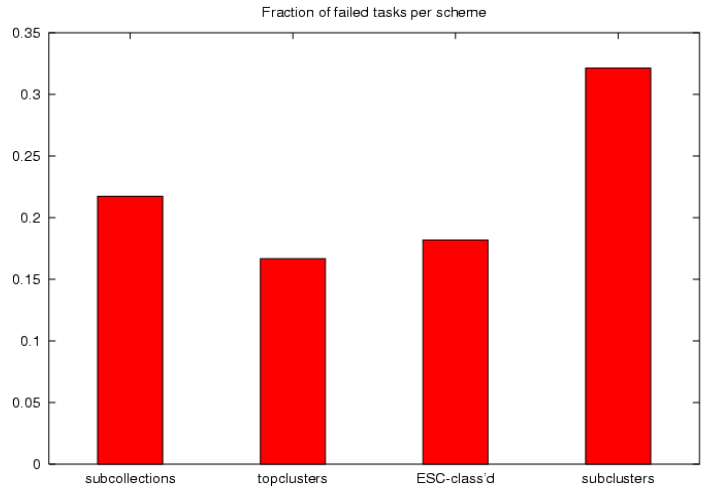
The resulting graphs function as a unique document browse scheme, allowing users to not only browse through a cluster hierarchy but also gain structural knowledge about makeup of an entire document corpus. We plan to include this visualized scheme in future browse experiments similar to those described above.



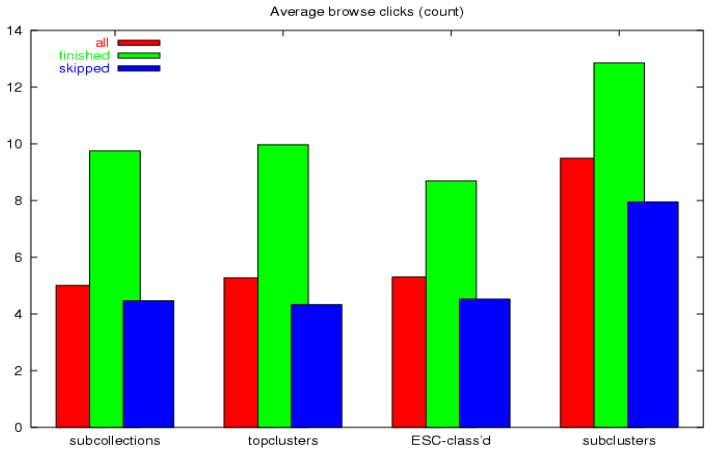
Two cluster visualizations. If a cluster contains a plus-sign then it contains subclusters. The left layout was done with PCA, while the right was done using a radial distance-based layout.

FINDINGS

Our results so far show that AI-derived schemes are competitive with the trivial mode of browse organization of harvested resources, and can add value to digital library browse services. Depending on the subject being researched, the user, and the scheme, the effectiveness of AI schemes may vary from poor to excellent. This variation is one of our key findings. We suggest that this means it is best to give users multiple alternative browsing options, using both the trivial and a variety of AI-derived schemes.



Fraction of failed browse tasks for each scheme. A "failed" (or "skipped") task is one where the user gave up searching for the resource, after an initial wait period.



Count of average navigation clicks needed to complete browse tasks, for all, finished, and skipped tasks. Note that despite being the only two-level scheme, subclusters does not require twice as many clicks.

FUTURE WORK

We expect that improvements in the tuning of our algorithms and our text processing methodologies will lead to improved accuracy and hence browse schemes that are more useful to end users. We plan on conducting experiments with fully-hierarchical clustering, and clustering over collections of web resources acquired via focused crawl, and combined collections of DL and web resources. These advances and evaluations will be conducted in the second half of the project timeline.

SOFTWARE

MetaCombine is producing free/open source software to carry out our evaluations and build demonstration digital library systems. Visit http://www.metacombine.org/ to see what is available.