Collaborative Classification of Growing Collections with Evolving Facets

Harris Wu
Old Dominion University
Norfolk, VA 23508, USA
1-757-683-4460
hwu@odu.edu

Mohammad Zubair
Old Dominion University
Norfolk, VA 23508, USA
1-757-683-5000
zubair@cs.odu.edu

Kurt Maly
Old Dominion University
Norfolk, VA 23508, USA
1-757-683-4817
maly@cs.odu.edu

ABSTRACT
There is a lack of tools for exploring large non-textual collections. One challenge is the manual effort required to add metadata to these collections. In this paper, we propose an architecture that enables users to collaboratively build a faceted classification for a large, growing collection. Besides a novel wiki-like classification interface, the proposed architecture includes automated document classification and facet schema enrichment techniques. We have implemented a prototype for the American Political History multimedia collection from usa.gov.

Categories and Subject Descriptors
H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces – *Collaborative Computing.*

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and indexing – *indexing methods.*

General Terms
Algorithms, Design, Experimentation, Human Factors

Keywords
Collaborative classification, faceted classification, social classification, tagging, wiki

1. INTRODUCTION
The rapid growth of the Internet along with advancements in content creation tools have resulted in large, growing collections with a variety of multimedia documents. It is difficult to explore a large collection without a classification scheme, especially non-textual collections for which keyword-based search has limited value. It is common to have a few experts to classify documents in a collection. Such a centralized approach for a large growing collection is prohibitively expensive and imposes a biased, static, rigid structure. Social tagging systems such as del.icio.us (http://del.icio.us/tag/wiki) and flickr.com, on the other hand, allow individuals to assign free-form keywords (tags) to any documents in a collection. While free, open and evolving, social tagging systems suffer from low quality and the lack of structure in tags. Ambiguity and noise arise from the linguistic nature of tags such as polysemy, homonymy, plurals, and synonymy.

HT'07, September 10-12, 2007, Manchester, United Kingdom.

We are developing a system that improves access to a large, growing collection by supporting users collaboratively to build a faceted classification. Different from a single hierarchy, a faceted classification allows assignment of multiple classifications to an object, supporting multiple user perspectives in exploration and making it suitable for collaborative approaches. Compared to searches, faceted classification allows users retrieve information through recognition of category names instead of recall of query keywords. Faceted classification has been shown to be effective for exploration and discovery in large collections [1]. Faceted classification consists of two components: the facet schema containing facets and categories, and the association between each document and the categories in the facet schema. For collections growing in both volume and variety, a major challenge is to evolve the classification: continuously classify new objects, modify the facet schema, and reclassify existing objects into the modified facet schema. Existing classification systems are typically centrally managed and difficult to evolve. Our system allows users to collectively 1) evolve a schema with facets and categories, and 2) to classify documents into this schema. Through users' manual efforts and aided by the system's automated efforts, a faceted classification evolves with the growing collection, the growing user base, and the shifting user interests. Our fundamental belief is that a diverse, large group of people can do better than a small team of experts.

2. RELATED RESEARCH
Our research is at the intersection of knowledge management, ontology generation, document classification, metadata management, collaborative systems, data mining, information retrieval and hyperlink navigation. Our approach resembles popular wiki and social tagging systems. Below we discuss several research projects closest to ours in spirit.

The Facetag project [2] guides users' tagging by presenting a predetermined facet schema to users. The system guides users' classification through a thoughtfully designed interface. However, predetermined facet schemas are insufficient in supporting user needs. Another problem is that user efforts alone cannot "adequately" classify documents into all pertinent facets. Wu [3] algorithmically creates a single global hierarchy from overlapping hierarchical classifications manually created by users. However

users may not be willing to create their own classifications due to the extensive efforts required. Also the resulting single hierarchy cannot accommodate diverse user perspectives. A recent trend is to create a classification schema from existing tags in social tagging systems. For example, [4] builds a hierarchy of tags from the data in Flickr.com using graph centrality analysis. Since the categories in these hierarchies are tags (freeform keywords), the categories inherit the problems with tags such as synonymy and polysemy. The automatic classification approach also suffers quality problems as with any data mining methods.

To our knowledge, no other researchers are building a system that utilize automated techniques to assist users build a faceted classification, both to build the schema and to classify documents into it, in a collaborative and interactive manner.

3. SYSTEM OVERVIEW

Figure 1 is a functional overview of the system. Utilizing the metadata created by users' tagging efforts and harvested from other sources, the system aids users to maintain an up-to-date classification of items in the collection. We focus on three novel features: 1) to allow users collaboratively build and maintain a faceted classification, 2) to systematically enrich the user-created facet schema, 3) to automatically classify documents into the evolving facet schema.

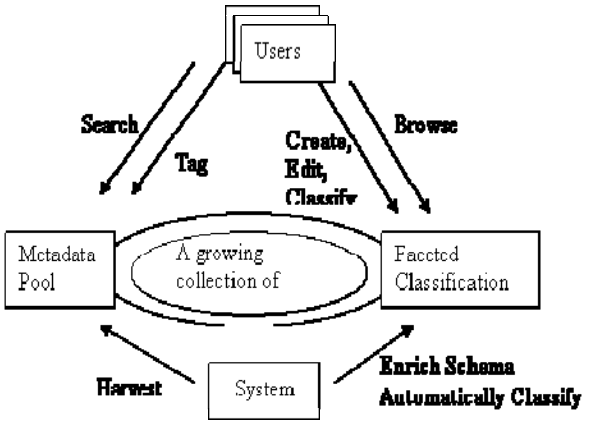

Figure 1. System Overview

We have developed a javascript-based interface that allows users create and edit facets/categories in a wiki fashion. Users can classify (or re-classify) documents by clicking and dragging documents into faceted categories. For automatic classification, we use a support vector machine method [5] utilizing users' manual classification as training input. For systematic facet enrichment, we are exploring ways to create new faceted categories from free-form tags based on a statistical co-occurrence model [6] and also WordNet [7]. Figure 2 shows the major system components, with novel features in bolded text. Next section discusses the implementation details in the context of prototype deployment.

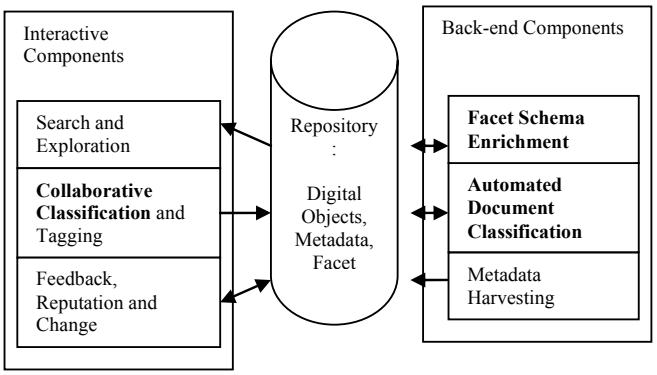

Figure 2. System Components.

4. PROTOTYPE IMPLEMENTATION

We have implemented a prototype on the American Political History (APH) image sub-collection of the US Government Multimedia Collection, a federated collection of millions of images, videos and other multimedia documents (http://www.usa.gov/Topics/Graphics.shtml). All the images in this APH collection are non-copyrighted, and therefore we can copy the whole collection. The APH collection currently contains 542 images, many of which are among the nation's most valuable historical documents. Currently users can browse either by era, such as 18th century and 19th century, or by special topics, such as "presidents" (Figure 3). There are only four special topics manually maintained by the collection administrator, which do not include most items in the collection. The description for each image is very brief and contains limited information. The collection lacks tools and metadata for users to explore and utilize.

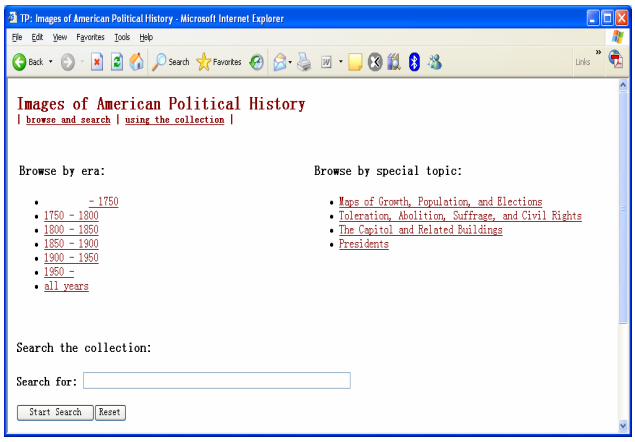

Figure 3. The American Political History Collection at usa.gov

After deploying our prototype, the collection now can be browsed through several facets (users can create more facets as needed) (Figure 4):

- Artifact (articles, maps, pictures, etc.)
- Location (New Jersey, New York, etc.)
- Foreign Fairs (England, France, etc.)
- Topics (buildings, presidents, etc.)
- Year (the existing Era facet: 18th century, etc.)

Figure 4. New Faceted Browsing Page for the Collection

Users can browse the collection using a combination of these facets. For example, a user can browse the "Maps" category, which includes a list of maps. Refining the list to "Washington, D.C" will reduce the list (Figure 5). More importantly, the collection now allows users to add structured metadata in a collaborative fashion. Next to each item there is a "Classify" button, which allows users to classify (or re-classify) the item (Figure 5).

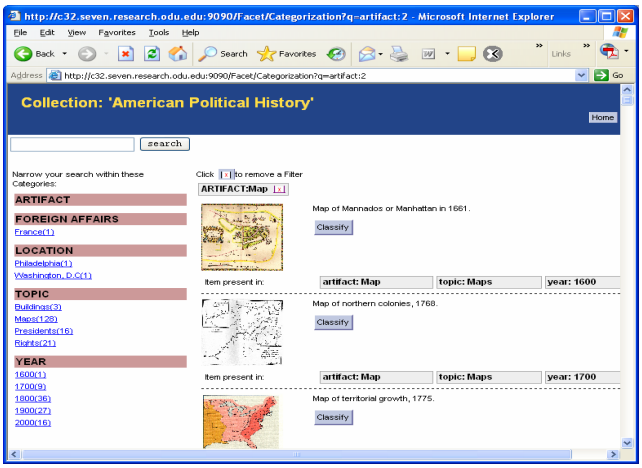

Figure 5. Navigating the facets

Users can classify the item using a drag-and-drop fashion, by dragging the image (or object icon) into the classification hierarchy, which will expand and collapse automatically as the object is dragged over (Figure 6). Users can also create or modify categories just as they manage directories in a file explorer.

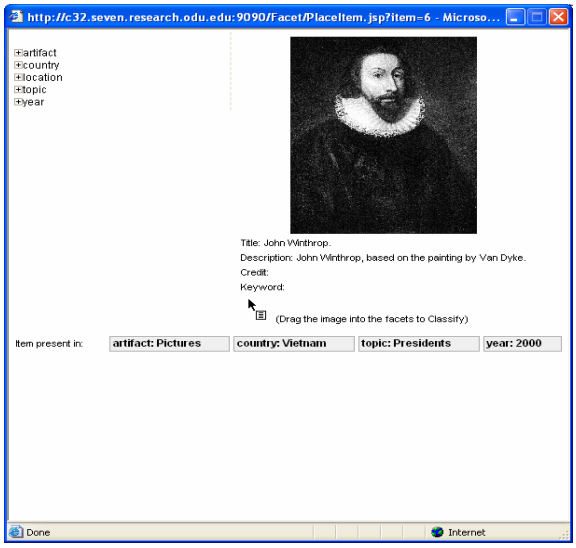

Figure 6. The Classification Interface.

Users' manual efforts and system's automated algorithms together construct and maintain a faceted classification of the collection. The automated algorithms will be discussed in the next section. Our system can be deployed on any existing large, growing collections, using the process illustrated in Figure 7.

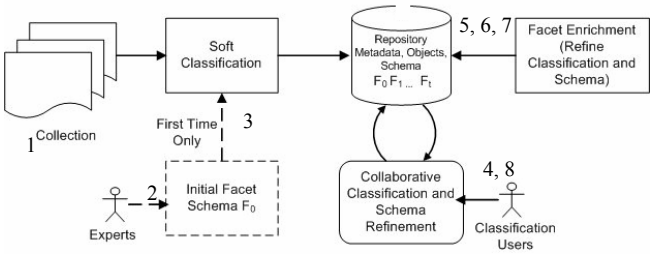

Figure 7. Deployment Process

The process (Figure 7) works as follows. 1) A collection C has *n* multimedia documents with no existing faceted classification. 2) An initial facet schema, F_0, is manually developed by a panel of experts, utilizing the metadata harvested from the collection. 3) A portion of the collection, T, is manually classified into F_0 by a panel of experts. The system tentatively classifies the rest of the collection into F using automated techniques, using T as a training set and harvested metadata (also content of textual documents) as input. The system-made classifications are called "soft" classifications. The soft classifications will be available to user search and exploration, but displayed in a grayed color. Upon users' approval, by explicit approval actions or providing positive feedback, a "soft" classification can be "hardened", i.e. officially added to the classification. The users can easily reject or modify the "soft" classifications if users find them inappropriate. 4) Over time, users continue to classify documents into the faceted schema. Users collaboratively edit the initial faceted schema, by add new facets, categories, or modify the existing ones. The schema evolves over time: F_0, F_1, …, F_t. 5) The system tentatively enriches the faceted schema, such as to add "soft" categories to an existing facet. 6) As the facet schema changes, the system will tentatively classify existing documents to new categories that belong to F'_t, a "stable" subset of the revised faceted schema F_t.

Again, the system-made classifications are "soft" classifications. Users will accept or reject these soft classifications as they use the system. 7) As new documents are added to the collection, the system will tentatively classify the document if any metadata is available. 8) As new users come to the collection, new documents are added to the collection, and as users' interests change, users will bring in new perspectives and evolve the facet schema.

5. SERVER-SIDE ALGORITHMS

Server-side algorithms complement users' manual efforts in maintaining a faceted classification. The prototype contains an automatic classification algorithm, which classifies all documents in the collection during initial deployment and also when new documents are added to the collection. Currently all server-side algorithms run as off-line programs, but they can be run periodically to address changes in the collection and update the schema and current classification.

For each category in the faceted classification schema, there is a SVM (support vector machine)-based classifier. Users' manual classifications are utilized as training input, and all keywords in the metadata record for a given document are used as the feature set. As users continue to classify documents or affirm system-generated classifications, the classifiers are regenerated periodically using enlarged training sets. SVM has been shown to be very effective in document classification [5].

We are experimenting with a facet enrichment approach that adds categories utilizing the metadata pool, a semantic lexicon (WordNet [7]), and a statistical co-occurrence model [5]. WordNet records synonyms and hierarchical relationships among words such as:

```
dog, domestic dog, Canis familiaris
  => canine, canid
    => carnivore
      => placental, placental mammal, eutherian, eutherian mammal
        => mammal
          => vertebrate, craniate
            => chordate
              => animal, animate being, beast, brute, creature, fauna
                => ...
```

Figure 8. Synonyms and Hypernym Hierarchies in WordNet

While WordNet is a powerful database, it focuses on the generic English language and does not capture all hierarchical relationships among concepts in a particular domain. The statistical co-occurrence model [5] identifies parent-child relationship between x and y if all documents tagged with y are also tagged with x (so-called subsumption). For example, if all images labeled with "liberty bell" are also labeled with "independence", whereas "independence" is an existing category, the algorithm will suggest "liberty bell" as a subcategory under the "independence" category.

For an existing facet f, existing category c, the algorithm identifies the set of keywords S_c in the metadata pool (i.e. the keywords that have been assigned to documents in the collection) that are siblings or children of c. The facet f is utilized for disambiguation when c has multiple meanings. The algorithm recommends these siblings or hyponyms of c as new categories parallel to c or as subcategories under c. The documents assigned with these keywords in S_c are then placed under the new categories.

6. EVALUATION AND NEXT STEPS

Our preliminary evaluation suggested two major items for improvement: 1) to prune the redundant (synonymous or overlapping) categories created by users; 2) to improve the accuracy of automated classification. For the first item, we are enhancing the facet enrichment algorithm so that it not only adds new categories, but also identifies and merges redundant categories. For the second item, we are exploring ensemble methods using a combination of classifiers for automated classification. We are making continuous improvements to the user interface, facet enrichment and automated classification techniques.

We are developing two new components for the system. One is a real-time component that classifies a new document into existing facets. When a user adds a document to the repository, the user interface can present the classification results and guide the user in classifying the document. This real-time technique will encourage users to prune and enter the metadata. The other component is to allow tagging, and to create facets out of free-form keywords, or tags. We are evaluating existing taxonomy generation techniques such as [4] and [6].

7. REFERENCES

[1] Hearst, M.A., Clustering versus Faceted Categories for Information Exploration. Communications of the ACM, 2006, 49(4).

[2] Quintarelli, E., L. Rosati, and Resmini, A. Facetag: Integrating Bottom-up and Top-down Classification in a Social Tagging System. EuroIA 2006, Berlin.

[3] Wu, H. and M.D. Gordon, Collaborative filing in a document repository. SIGIR 2004: p. 518-519

[4] Heymann, P. and Garcia-Molina, H., Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Stanford Technical Report InfoLab 2006-10, 2006.

[5] Joachims, T. Text categorization with support vector machines. In Proceedings of 10th European Conference on Machine Learning, pages 137–142, April 1998.

[6] Sanderson, M. and B. Croft, Deriving concept hierarchies from text. SIGIR 1999: p. 206-213.

[7] WordNet: An Electronic Lexical Database. Christiane Fellbaum (editor). 1998. The MIT Press, Cambridge, MA.