Building Dynamic Image Collections from Internet

Fu, Liuliu
luna.foe@gmail.com
Old Dominion University, USA

Maly, Kurt
Old Dominion University, USA

Wu, Harris
hwu@odu.edu
Old Dominion University, USA

Zubair, Mohammad
Old Dominion University, USA

People often want to collect and utilize free, publicly available images on a given subject. Image sharing systems such as Flickr store billions of user-contributed images. While such systems are designed to encourage user contributions and sharing, they are not well-organized collections on any given subject. We propose an approach that systematically harvest images from Internet and organize the images into an evolving faceted classification. We implemented a prototype to continuously harvest the most popular images on Flickr related to African American history, and organize them into an evolving faceted classification collaboratively maintained by users. The same approach can be applied to other digital humanities resources on the Internet. The talk will elaborate the details of technical design and prototype implementation, and discuss evaluation results.

2. Introduction

Flickr hosted over 4 billion images as of October 2009 and is growing by about 4 million pictures a day. Facebook hosted 15 billion photos and Imageshack hosted 20 billion images as of April 2009. Other photo sharing sites such as Picasa, Multiply and PhotoBucket also host billions of images [Schonfeld, 2009]. In contrast, organized public domain image collections are relatively scarce [Wikipedia: 'Public domain image resources']. The largest public domain image repository, Wikimedia Commons, reached 5 million images as of September 2009.

It would greatly benefit the humanities community if images in those image sharing sites can be organized and utilized. We propose an approach that systematically searches and harvests images (actually the link to the image and metadata, but not image itself) from image sharing sites, and organizes the images into a multi-faceted classification. The data harvesting is performed on a continual basis, and the classification evolves over time. Besides automated programs, the approach utilizes collaborative human efforts to improve the quality of collection. We implemented a prototype that builds a dynamic image collection on African American History from the most popular images on this subject on Flickr.com. Our fundamental belief is that a large, diverse group of people (students, teachers, etc.) can do better than a small team of librarians or editors in constructing a multimedia collection at virtually no cost.

Note that not all the images on those sharing sites are copyright-free or have a creative commons license. However, most of the sites allow other websites to directly link to their images if the images are marked as public access by their contributors, and if credits are properly given. Our approach displays images through embedded image URLs but does not download the images from their original sources.

3. Related Work

Many are trying to utilize the images in fast-growing photo sharing and social networking sites. For example Getty Images, the leader in stock photography, hires image editors to select most popular Flickr images and obtain copyright from individual contributors, then sells the images for $5 per image (`http://www.gettyimages.com/`). Computer-graphics researchers at the University of Washington have utilized Web images to digitally reconstruct buildings in 3-D. For example, based on 150,000 publicly accessible Flickr pictures of Rome, the program automatically re-created the Colosseum, Trevi Fountain, and the outside and inside of St. Peter's Basilica, among other Roman icons. The technique can be used to make virtual-reality experiences for tourism, auto-build cities for video games and movies, or help digitally preserve and study historic cities that are being destroyed by human-caused or environmental factors [Jaggard 2009].

Researchers have argued for building an academic Flickr, or an academic photo sharing site in general: a net-based service that would enable faculty and researchers to post and share images with scholarly value, either with the general community, or pursuant to any associated rights, to restricted-use populations [P. Brantley's blog]. For example, a group at Lewis & Clark College in Portland is in the process of developing an educational collection of contemporary ceramics images using the photo sharing site Flickr [McWilliams 2008].

Our project attempts to build free, well-organized topical images collections from the images contributed by Internet users, to support education or research objectives. While most photo sharing systems support keyword-based search utilizing user-contributed metadata, none of them support browsable hierarchies that allow users to explore a given subject in depth. Using librarians or images editors to manually construct a topical collection is cost prohibitive, and unfeasible if the collection needs to keep up with rapidly growing data sources. Our collection-construction approach combines the collaborative concepts of wiki and social tagging systems with automated classification techniques. Our system allows users to collaboratively build a classification schema with multi-faceted categories, and to classify documents into this schema. Utilizing users' manual efforts as training data, the system's automated techniques build a faceted classification that evolves with the growing collection, the expanding user base, and the shifting user interests.

4. Architecture and Prototype Implementation

Our collection construction approach is summarized in Figure 1. The system first collects images (links and metadata such as tags) on a given topic using keyword search, utilizing the APIs (Application Programming Interface) provided by image sharing sites or search engines. For the initial collection, a group of experts or administrators create the initial classification schema and classify a set of images into the initial schema. Utilizing experts' classifications as training data, and also Wordnet and Wikipedia as knowledge bases, the system employs automated techniques (heuristic matching rules and support vector machine-based classifiers) to classify images into the classification schema. In a wiki fashion, users of the image collection can modify and improve the classification schema, and manually classify items into the schema. Users can also assign additional tag or annotations to image objects. Utilizing the additional metadata from users' tagging and annotation efforts and by analyzing users' classification/usage history, the system refines both the classification schema and the item-category associations. The system continues to collect and classify images to stay up-to-date with external image sources.

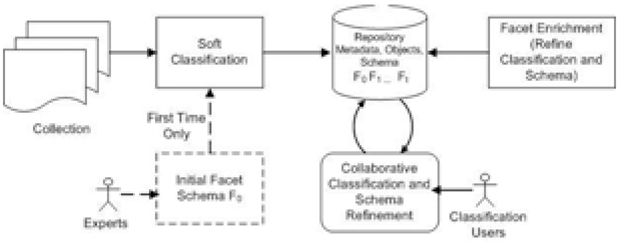
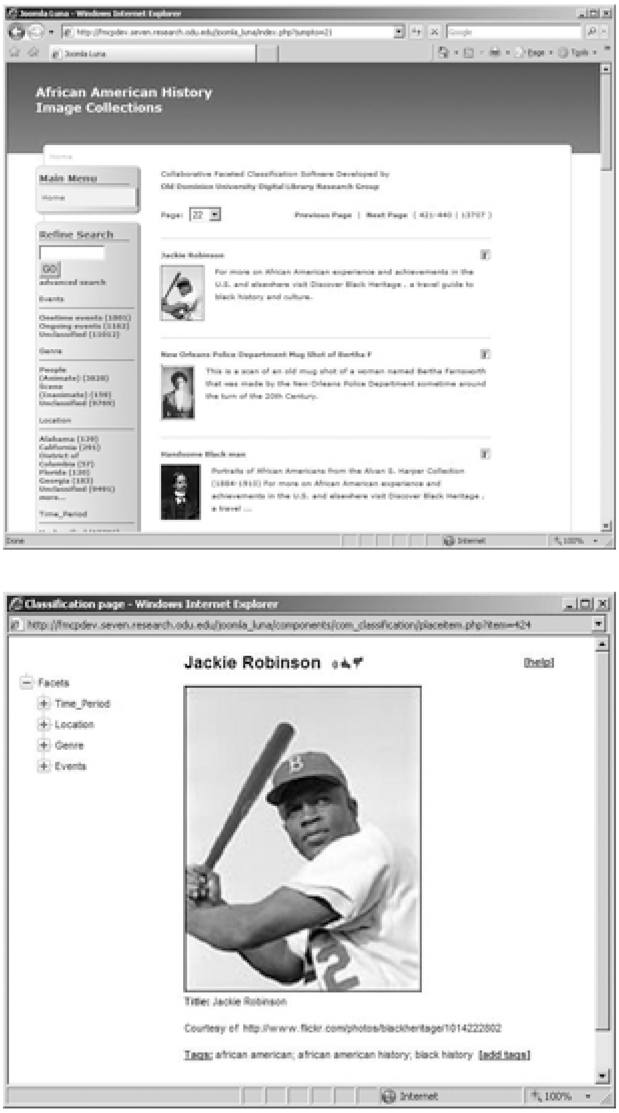

Collection
Soft Classification
First Time Only
Initial Facet Schema F_0
Experts
Repository Metadata, Objects, Schema $F_0 F_1 ... F_i$
Facet Enrichment (Refine Classification and Schema)
Collaborative Classification and Schema Refinement
Classification Users

Figure 1. Systematic approach of constructing a topical collection using Internet images.

We built a prototype to construct an image collection on African American History from Flickr. By querying "African American History" in the search field, we extracted metadata for all the images in the result pages: title, url, description, tags, and the contributor. The initial collection contained about 11,000 Flickr images on African American History. Over 3 months the collection has grown to contain about 13,000 images. During the conference we will elaborate the details of technical design, prototype implementation, and the evaluation results. Figure 2 shows the browsing and classification interfaces of our prototype.



Joomla Luna - Windows Internet Explorer

African American History Image Collections

Collaborative Faceted Classification Software Developed by Old Dominion University Digital Library Research Group

Page: 22 Previous Page | Next Page (421-440 | 13707)

Jackie Robinson

For more on African American experience and achievements in the U.S. and elsewhere visit Discover Black Heritage , a travel guide to black history and culture.

New Orleans Police Department Mug Shot of Bertha F

This is a scan of an old mug shot of a woman named Bertha Farnsworth that was made by the New Orleans Police Department sometime around the turn of the 20th Century.

Handsome Black man

Portraits of African Americans from the Alvan S. Harper Collection (1884-1910) For more on African American experience and achievements in the U.S. and elsewhere visit Discover Black Heritage , a travel ...

Classification page - Windows Internet Explorer

http://fmcpdev.seven.research.odu.edu/joomla_luna/components/com_classification/placeitem.php?item=424

Jackie Robinson

Title: Jackie Robinson

Courtesy of http://www.flickr.com/photos/blackheritage/1014222802

Tags: african american; african american history; black history [add tags]

Figure 2. Browsing and classification interface of the prototype.

5. Discussion

Evaluation of the prototype in a classroom environment shows promise. Measured by metrics such as precision, recall and image quality (popularity), the prototype is more effective than Flickr in supporting several image retrieval tasks. The evolution of classification shows improvements, based on user ratings of categories and category-item associations. We conducted interviews and usage observations, which help understand the level of efforts that users spend on tagging and classification. For future research, we are interested in whether social tagging and tag convergence [Muller *et al.* 2008] can be utilized to assist or substitute classification efforts.

Our approach can be applied to other digital humanity resources. For example, we have developed another prototype to construct a dynamic collection of news items on a given topic based on Google News. We believe that a combination of collaborative and automated classification techniques can construct valuable digital humanities collections at low costs.

As far as we know, no one has combined user efforts and automated techniques to build a faceted classification. Several research projects are related to social tagging and classification, however. Several projects attempt to construct tag hierarchies or ontologies, or otherwise harvest the intelligence stored in tags [Heymann and Garcai-Molina 2006, Schmitz and Patrick 2006, Harris *et al.* 2006]. Our earlier work [Arnaout *et al.* 2008] on faceted classification was presented in the Digital Humanities 2008 conference.

6. Acknowledgements

This project is supported in part by the United States National Science Foundation, Award No. 0713290.

References

Schonfeld, Erick (2009). 'Who Has The Most Photos Of Them All?'. *TechCrunch*. April 7, 2009. `http://www.techcrunch.com/2009/04/07/who-has-the-most-photos-of-them-all-hint-it-is-not-facebook/`.

Wikipedia:Public Domain Image Resources. `http://en.wikipedia.org/wiki/Public_domain_image_resources`.

Jaggard, Victoria (2009). 'Flickr Pictures Help Build 3-D Rome in a Day'. *National Geographic News*. September 24, 2009.

Jaggard, Victoria (September 24, 2009). *Flickr Pictures Help Build 3-D Rome in a Day. National Geographic News*. `http://news.nationalgeographic.com/news/2009/09/090924-flickr-rome-build-day.html`.

Brantley, Peter. *Design Beyond the Interface* Blog `http://blogs.lib.berkeley.edu/shimenawa.php/2008/04/17/ah_screw_the_interface`.

McWilliams, Jeremy (2008). 'Developing an Academic Image Collection with Flickr'. *Code{4}lib Journal.* **3**.

Muller, M.J., Dugan, C., Millen, D.R. (2008). 'Metrics for sensemaking in enterprise tag management'. *CHI 2008 Sensemaking Workshop.* Florence, Italy, April 05 - 10, 2008, pp. 1493-1496.

Heymann, P., Garcia-Molina, H. (2006). *Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems.* University of Southampton: Stanford Technical Report InfoLab. `http://ilpubs.stanford.edu:8090/775`.

Schmitz, Patrick (2006). 'Inducing Ontology from Flickr Tags'. *World Wide Web Conference 2006 (WWW2006).* Edinburgh, UK., May 22–26, 2006. Collaborative Web Tagging Workshop. .

Wu, Harris, Maly, Kurt, Zubair, Mohammad (2006). 'Harvesting social knowledge from folksonomies'. *Hypertext 2006 — Seventeenth ACM Conference on Hypertext and Hypermedia.* Odense, Denmark, 23-25 August 2006.

Arnaout, Georges, Maly, Kurt, Mektesheva, Milena, Wu, Harris, Zubair, Mohammad (2008). 'Exploring Historical Image Collections with Collaborative Faceted Classification'. *Digital Humanities 2008.* Oulu, Finland, June 25-29, 2008.