Finding Your Inner Metadata

Using machine learning to discover patterns in large image collections

Aaron Coburn, Amherst College
acoburn@amherst.edu

For repositories of any size, quality metadata is of utmost importance. Without it, users cannot find objects and librarians cannot maintain their collections. The better the metadata, the easier it is to work with these collections. Sometimes, however, the metadata is either not available or so minimal to be of little practical value. Still, it may be clear that such collections are of significant value to an institution. And while for small collections, one could manually enhance object metadata, at larger scales this becomes distinctly infeasible.

Historically, at Amherst College, the slide negatives for all photographs were archived by the library. For nearly two hundred years, these negatives were organized and preserved; anyone looking for a photo from the fall of 1954 can easily find it. In 2004, however, the college photographers stopped using film and began recording events with digital cameras. Along with the change to digital photography, the volume of generated material increased dramatically: instead of shooting one or two rolls of film at an event, a typical event may now result in thousands of files. In the intervening decade, not a single one of those photos made its way to the archives. Instead, the photos are stored on a variety of network file systems, loosely organized in folders according to somewhat idiosyncratic conventions.

Nearly a million photos later, we have begun a project not only to transfer those digital images into a Fedora4-based repository system, but also to figure out what this collection contains. An estimated 10% of the photos have useful embedded metadata, such as location, event and other descriptive keywords. While these keywords do not use a standard controlled vocabulary, they are accurate and, for the most part, consistent. Still, much of this will need to be moved into standard controlled vocabularies. For the remaining files, the only reliable metadata is the creation date and the directory location where it was originally stored.

Archiving and processing this quantity of material is not entirely trivial. The raw files comprise nearly 25 Terabytes of data. And simply to be able to handle that quantity of data within a reasonable timeframe, the architecture of the system must be distributed horizontally across an elastic array of servers.

As described above, there will also be several distinct categories of metadata. Timestamps, technical metadata and any ID3 tags added by the college photographer can, generally, be trusted at face value. For much of the collection,

however, we will be using an automated process of metadata enhancement in which the generated data may be suspect. In particular, we are using a collection of machine learning algorithms in an attempt to assert relationships between similar images. While that metadata will be useful, it is of a much different character, and we will need to be able to easily partition that data into separate systems.

The RDF model provided by Fedora4 allows us to model these different categories of assertions along with a corresponding degrees of certainty. We are also able to extract that data selectively into external systems for refinement of this relationship and similarity model.

For the image processing, we are using a set of stochastic algorithms to identify similarity. For all of these, it is infeasable to construct a complete pairwise similarity matrix: such a matrix would contain nearly 1 trillion entries. Instead, we are using a Markov chain Monte Carlo simulation to create a minimal spanning tree across the collection, represented as a sparse graph. This representation will make it reasonably easy to generate clusters of similar images. The image processing, itself, uses some standard techniques (particularly with the application of Laplacian and Gaussian filters) to identify scale- and orientation-invariant patterns in the images themselves. By representing each image as as vectors of “feature locations”, they can be efficiently compared and a similarity measure can be efficiently and reliably calculated.

Thus, by bootstrapping from the roughly ten percent of the collection with trustworthy metadata, it is our hope that similar images can be discovered from within the 90% of untagged images. This process will impose a layer of structure, albeit tentative, on an otherwise unstructured collection. Then, after passing through an interface where a librarian can validate the automated classification, the known metadata is assigned to more “trustworthy” RDF-based properties.

The overall architecture of this includes Fedora4 as a central repository. Some ancillary indexing systems, such as Solr and Fuseki will be used, as will a cluster of image processing machines. Much of the orchestration for this processing will be handled asynchonously by a set of OSGi-based services. The scale of the project also lends itself well to a system such as Fedora, which is capable of supporting millions of objects and billions of RDF triples.

The end goal for the project is twofold: first, to have the image corpus in a reliable system capable of being preserved and managed by the college archivists, and second, to have accurate metadata on as much of the image collection as possible. By June, 2015, the major components of this will be complete and we will report on how well it worked, and how a Fedora-based infrastructure was able to handle collections of this size.