University of Southern California

Directed Research Report

2014-12-12

Mengying Wang

Prof. Chris Mattmann

USC ID: 4641829269

Project Background

The Arctic region is one of the Earth's last great frontiers [1]. The region is home to strong national security interests (e.g., land/air/sea and maritime), commercial interests (e.g., in oil, and in minerals and natural resources), and is heavily influenced by the Earth's changing climate including temperature increases which have led to urgent problems including glacial melting [2]. Several strategic priorities for the region are identified in the U.S. President's National Strategy for the Arctic Region document [1], including *Protect the Arctic Environment and Conserve Natural Resources*, etc.

Although myriad data sources relevant to the Arctic region are already available to increase understanding, the challenge remains sifting and sorting through those data (we refer to this process as "data triage") to properly respond to the regional priorities laid out in the U.S. national strategy. During this "data triage" process, there are a variety of data formats that a researcher must parse, analyze and interpret, not to mention the variety and quality of associated metadata available that a researcher must make sense of. For example, in the NASA's Antarctic Master Directory (AMD) [3] system, of the 50 data results, 52% have an Unknown data format; the majority of the known data types (16%) are ASCII (text) format; the remaining are PDF files (11%), Excel files (6%), scientific data formats (NetCDF [13] – 3%), movie files (MP4 – 2%), Microsoft Word files (2%) and HTML files (2%).

This project will construct and deliver a classification framework based on the open source Apache Tika technology [4] that obviates the hetereogeneity of Polar data formats, and that provides a unified mechanism for analyzing and interpreting the metadata and data once classified. Specifically, the research will comprise three steps:

1. Classify and understand Polar and Arctic data and metadata.

2. Augment Apache Tika to automatically detect the most prevalent Polar and Arctic data and metadata formats.

3. Augment Apache Tika to automatically parse the most prevalent Polar and Arctic data and metadata formats.

Job duties

I have been working on the first step - the classification and triage of the Polar and Arctic data. I need to deploy Apache Tika at scale using Apache OODT [5, 6], Apache Nutch [7]

and Apache Solr [8]. Apache OODT is a data management and processing framework originally developed at NASA; Nutch is a distributed search engine that runs on top of Apache Hadoop, and that indexes content metadata (including type, and language information generated by Tika). Both OODT and Nutch can index metadata in the Solr indexing engine. Once MIME types are indexed in Solr, we can easily perform MIME type classification and develop data format distributions and profiles for different representative Polar data sources, e.g., NASA AMD, PolarGrid [9], Acadis [10], etc. The metadata distribution will characterize the number and heterogeneity of Polar data formats from these representative data sources, and inform the next step.

After three months' hard work, I have successfully installed the Apache OODT, Apache Nutch, Apache Solr and Apache Tika on my Mac. To crawl and index Polar websites, I have applied two approaches:

1: Apache OODT, Solr and Tika Integration:

First, employed the Apache OODT Pushpull component to crawl a bunch of data files from the remote ftp servers to the localhost. Second, with the help of the Apache OODT Crawler, I could automatically call the Apache Tika to extract the metadata information for each data file, and then both the data and metadata files would be ingested into the Apache OODT File Manager. Third, used the SolrIndexer tool to dump all data as well as their metadata from the Apache OODT File Manager to the Apache Solr.

2: Apache Nutch, Solr and Tika Integration:

First, configured the Apache Solr to be integrated with the Apache Nutch. Second, used the Apache Nutch to crawl a number of webpages, delegating indexing and searching to the Apache Solr.

Although both OODT and Nutch could index metadata in the Solr, they may have their own strengths and weaknesses. In my opinion, the OODT has so many useful components, e.g., CAS-Filemgr, CAS-Workflow, CAS-PGE, CAS-Crawler, etc. They could be flexibly combined together to complete a number of complex tasks. However, to combine these components together, it is really hard and time-consuming for a novice. It would be much better to have something like RADiX OODT, which is an effort to build a distribution of OODT that sets up, installs, and runs within a few commands. By contrast, the Nutch is a well matured, production ready web crawler. Only a few configurations, you could start crawling the whole websites using a single command with few runtime errors. However, the Nutch automatically stores the content to the Hadoop filesystem, making it more scalable and efficient but otherwise requiring specialized filesystem APIs to retrieve; whereas the OODT stores the content to local storage as native files in a less efficient and scalable, but more easily accessible fashion.

As I have said before, since the Apache Nutch is relatively easy to use, I used the second approach: Apache Nutch, Solr and Tika Integration to crawl and index Polar data sources: NSF ACADIS and NASA AMD. For the former, overall 2929 documents are crawled, with 6 content types, including application/xhtml+xml (2374), application/zip(554), application/vnd.openxmlformats-officedocument.wordprocessingml.document(1), application/pdf (1), image/jpeg (1), application/atom+xml(1). For the latter, there are 29642 documents in all, with 11 content types, including text/html (30166), application/xhtml+xml (1449), image/png (513), image/gif (515), application/rss+xml (69), text/css (34), application/javascript (29), application/atom+xml (22), image/jpeg (17), application/xml (13), and text/x-java-source (3). However, after carefully analyzing the crawling result, I found that a number of documents with various content types, e.g., *.xlsx, *.xls, *.doc, *.docx, *.csv, *.m, etc, are not crawled in my crawling process. It turned out that there were some "robots.txt" files in these two websites, which didn't allow the Nutch to crawl these scientific data. By modifying the Nutch source code to disable the robot checks in the Fetcher class, and recompiling it, the “Robot denied” errors could be solved completely. Moreover, in NASA AMD, a lot of data download urls are generated by dynamic javascript. Currently, the Nutch parse-js plugin doesn’t have a good support to the javascript, so I need to try the nutch-selenium plugin to fetch javascript pages in the future.

Experience gained

In this directed research project, I have leveraged a combination of several Apache software technologies, including OODT, Nutch, Tika, Solr, Hadoop, and HBase. I have made some contributions to the open source community by:

1. Creating Jiras and submitting patches:
   OODT-756, OODT-766, OODT-781, OODT-783, NUTCH-1884, NUTCH-1893, etc.
2. Updating Confluence pages:
   OODT Push Pull User Guide and OODT Push Pull Plugins.
3. Discussing problems in the dev email list

Also I have earned the precious opportunity to join the OODT project management committee to continue contributing something for the Apache open source community, which is really a valuable experience to work with so many talented engineers!

1. U.S. National Strategy for the Arctic Region, http://www.whitehouse.gov/sites/default/files/docs/nat_arctic_strategy.pdf, May 2013

2. Overpeck, Jonathan, et al. "Arctic environmental change of the last four centuries." Science 278.5341 (1997): 1251-1256.

3. Antarctic Master Directory – NASA, http://gcmd.nasa.gov/portals/amd/, Accessed July 2013.

4. Mattmann, C., and J. Zitting. Tika in Action. Manning Publications Co., 2011.

5. Mattmann, C. A., Crichton, D. J., Medvidovic, N., and J. S. Hughes. "A software architecture-based framework for highly distributed and data intensive scientific applications." In Proceedings of the 28th international conference on Software engineering, pp. 721-730. ACM, 2006.

6. Mattmann, C. A., et al. "A reusable process control system framework for the orbiting carbon observatory and NPP Sounder PEATE missions." Space Mission Challenges for Information Technology, 2009. SMC-IT 2009. Third IEEE International Conference on. IEEE, 2009.

7. Cafarella, M. and D. Cutting. Building. "Open source search." Queue. v2 i2 (2004): 54-61.

8. Apache Solr, http://lucene.apache.org/solr/, Access July 2013.

9. Guo, Z., Raminderjeet S., and M. Pierce. "Building the polargrid portal using web 2.0 and opensocial." Proceedings of the 5th Grid Computing Environments Workshop. ACM, 2009.

10. Jodha Khalsa, S., et al. "The Advanced Cooperative Arctic Data and Information Service (ACADIS)." EGU General Assembly Conference Abstracts. Vol. 15. 2013.