Directed Research Proposal

-Prasanth S Iyer
piyer@usc.edu

The idea

One of the main challenges in extracting Polar data from the NASA Antarctic Master Directory (AMD) datasets is that these files belong to many different MIME types and the MIME-classification algorithm currently used by Tika is not accurate enough to predict the correct MIME type of each file. A possible reason for this is that the metadata provided about the types of these files is often incorrect and results in Tika classifying them incorrectly. My idea is to explore a machine-learning based algorithm that would learn to classify these files initially into top-level media types such as text, image, audio etc. and then, into more specific subtypes like xml, html etc. This classification will be based on features like the file extension or the "magic" bytes at the start of the file or the presence and frequency of specific characters within the file.

My research plan

I plan to follow the following steps in building my algorithm:-

1. Run a benchmark using Tika and Nutch on the existing dataset to understand how many files Tika is currently able to classify and parse correctly. This will enable us to do 2 things:-
    a. Analyze the dataset to find the number of files with incorrect metadata information in AMD and recognize how this incorrect metadata affects Tika's classification of files.
    b. Compare the benefits of using a ML-based approach over the existing classification algorithm and decide if we could simply refine the existing classification algorithm to work for a large percentage of the files in the dataset.

2. Once we decide that a ML-based approach is indeed necessary, I would explore the usual parameters necessary in a ML-based algorithm:-
    a. **The training set:** The ML algorithm would require a well-labeled representative training set from the AMD datasets. I believe I can retrieve this using the Nutch crawler.
    b. **Feature extraction:** The algorithm would also require a set of features for it to learn. I can reuse some of the parameters used by Tika's existing classification algorithm like the file extension and the magic byte pattern. However, I may also need to detect other parameters such as the frequency of certain characters in the file content. I can extract this using cross-validation on the training set.
    c. **Classification algorithm:** The ML algorithm would classify files into different MIME types based on the features. In some cases, the ML algorithm may not be able to predict the MIME type of a file with a reasonable confidence level. Such files would be characterized as files with unknown MIME-types.
    d. **Unknown file formats:** Given a file of an unknown file format, the algorithm should at least be able to give a top-level classification of whether it is a text file or an audio/video file or an image.

e. **Test data set:** The algorithm will then be tested over the entire AMD cryosphere dataset to classify file types with a reasonable accuracy.

Conclusion

The main advantage of using the ML-based approach is that we can easily extend the algorithm to classify file types over other datasets like PolarGrid and Acadis. Even unknown file formats can be classified up to a reasonable top-level MIME type. However, it is important to know if the ML-based approach is an overkill given Tika's existing classification algorithm. For this, we need to measure the accuracy of the existing algorithm on the dataset and this makes the results of the benchmark important.