The NASA/IPAC Extragalactic Database (NED) routinely reviews journal articles to extract fundamental data for extragalactic objects and join them across the spectrum into the database.

The process of reading journal articles to determine if they have data appropriate for inclusion in NED, and to determine what types of data they contain, is very labor intensive, especially with the ever-increasing numbers of publications and the growth of data each year.

We present a machine learning approach developed recently to help automate the classification of journal articles and their data content. Application of this approach can reproduce the classifications of a human expert to an accuracy of over 90%.

**ABSTRACT**

The NASA/IPAC Extragalactic Database (NED) is a comprehensive database of multiwavelength data for extragalactic objects, providing a systematic, ongoing fusion of information integrated from hundreds of large sky surveys and tens of thousands of research publications.

To keep the database content current, NED routinely reviews, extracts, and integrates data from high-impact journals including A&A, AJ, ApJ, ApJS, and MNRAS. In 2018, NED reviewed 14,123 articles in total, and a third of them were classified as NED appropriate papers (NAP). Once the topic of an article is determined to be relevant, the different data types provided in the article are further classified. Traditionally, this process is done by human inspection, which means an average of ~270 articles need to be reviewed every week to keep the classifications up to date.

In 2019, as part of IPAC’s Technology Initiative, we began exploring the application of machine learning (ML) techniques to assist with automating the classification of journal articles.

**INTRODUCTION**

The algorithm we chose is a support vector machine, which finds a hyperplane in the N-dimensional space that best separates the given classes by maximizing the distance between the plane and the data points. The specific package we used is the Stanford Classifier developed by the Stanford Natural Language Processing Group. This is supervised learning, as we deploy priors from many years of human classifications on multiple journals.

Figure 1 illustrates the process for the most basic case, which is the classification of NAPs and NNAs (Non NED-Appropriate papers). We obtain the HTMLs of the published articles from the journal websites. We then extract features from each article in a training set, combine them, and feed them into the machine learner together with the labels of the articles. At the heart of this is a maximum entropy (SoftMax) classifier where weights are generated based on the labels. Once we have the model, we then feed it with the features from new articles and it will output a label for the article with a probability attached.

**METHOD**

To keep the database content current, NED routinely reviews, extracts, and integrates data from high-impact journals including A&A, AJ, ApJ, ApJS, and MNRAS. In 2018, NED reviewed 14,123 articles in total, and a third of them were classified as NED appropriate papers (NAP). Once the topic of an article is determined to be relevant, the different data types provided in the article are further classified. Traditionally, this process is done by human inspection, which means an average of ~270 articles need to be reviewed every week to keep the classifications up to date.

In 2019, as part of IPAC’s Technology Initiative, we began exploring the application of machine learning (ML) techniques to assist with automating the classification of journal articles.

**TOPICS CLASSIFICATION**

We began by training a ML classifier to identify if the topics of an article are NED appropriate or not (i.e., NAP vs NNA). We used the priors from 16 years (2003-2018) of AJ articles. The text corpus contains 6,987 articles, and 1,984 of them are hand-classified as NAPs. An overall 100 rounds of 10 folds of the Stanford Classifier were processed on all 1,984 NAPs, and 1,984 randomly selected articles from the NNAs category. Essentially, this created and tested 1,000 classifiers with different choices of input from among the training items. We took the default probability cutoff at 0.5 to label the NAPs and NNAs.

When testing the best-performed model on the AJ articles, we were able to reproduce the priors to an accuracy rate of 97% and a recall rate of 94%. When testing on other journals (A&A, ApJ, ApJS, and MNRAS) from the same period (2003-2018), we were able to reproduce the priors to an accuracy of >91% for all of them.

This ML classifier (v1.0) has been integrated into the NED production pipeline. To keep up with the new information from journal articles, we do hand classifications on one or two volumes per journal every 6 months, and aim at a retraining/re-evaluation frequency of once per year. We have recently completed hand classifications on a few 2020 journal volumes. When comparing with the ML output, we found that our v1.0 classifier is still performing well at 92% for both accuracy and recall.

**CONTENT CLASSIFICATION**

In 2020, we focused on applying ML to further identify the data content in each article. We started with astrometry, photometry and redshift data as these are among the most fundamental properties NED keeps for an object. We also included a check for a NED acknowledgement in each article. Table 1 provides an example of the priors in NED that we try to reproduce.

We assembled a training set containing 9 years (2011-2019) of articles from the core journals NED reviews. We then selected 90-95% of the positives for training, and reserved 5-10% for testing. For each of the five classes (NAP, acknowledgement, astrometry, photometry, and redshift), 20 classifiers were generated using randomly selected subsets. Table 2 presents the accuracy and recall rates for the best set of classifiers when evaluating on the testing set. We achieved an accuracy rate of >92% for all five categories. The recall rates are at >93% for all except the acknowledgement classifier, which is under further investigation.

**CONCLUSIONS**

NED has explored using machine learning techniques to help identify the topics of astrophysics journal articles and different types of data contained in the articles. We show that we can reproduce the NED historic classifications, and make new predictions to an accuracy of 92% or greater.

We are in the process of building a v2.0 classifier (with data type classifications) into the NED production pipeline. Once implemented, it will further reduce the time the production staff spend on this labor-intensive enterprise, and allow us to focus human expertise on tasks that have little or no hope of being handled with AI/ML anytime soon, such as analyzing and understanding the data content well enough to clean, reformat, apply or validate metadata to prepare input to the database, carrying out probabilistic cross-matching and fusion of data from catalogs, and ultimately, making scientific analysis and discoveries with the data in NED.

2. We generated the features by splitting the words on space, using 2-5 letters n-grams, either at the beginning or at the end of the words. A typical article generally renders on the order of 100,000 features.