Galaxy Data Mines

Comparing Astronomical Object Classifications in NED and SIMBAD

Matthew Shubat
CS4490Z Thesis
Department of Computer Science
Western University

October 6, 2018

Supervisor: Dr. Pauline Barmby
Course Instructor: Prof. Nazim Madhavji

1. Synopsis

Due to the fundamental nature of their work, which spans the scope and scale of the universe, astronomers are faced with a vast set of data—with the potential size of being unimaginably, staggeringly vast. As can be imagined thoughtfully organizing, as well as searching through, even a subset of this data is a substantial undertaking. There are countless repositories of astronomical literature and data available online: each with their own methodology of organization and access. Although fruitful in breadth and depth, this landscape provides challenges to researchers in terms of finding the information they need in order to answer the questions they want to ask.

To help mitigate this challenge, several large aggregate data repositories have been developed. Two substantial instances of these are NED and SIMBAD. These databases have been immensely helpful in addressing this information organization and access challenge and have helped birth a new era of astronomical discovery.

An issue which still remains in each of these resources is the challenge of classification. There are billions of objects which need to be classified: a task of this magnitude necessitates some automated means of doing so. Between two such systems some errors or disagreements inevitably occur, and this results in an object having a different classification in each system.

In this paper, we propose an automated solution to help mitigate this issue. Our system would create a common classification hierarchy to which objects from NED's 1-Dimensional classification system and SIMBAD's hierarchical system would map. After this mapping occurs an automated hierarchical comparison would then be done in order to determine the level of compatibility of the classifications between NED and SIMBAD. This cross comparison would provide insight into the probability of a classification agreement occurring, further increasing the confidence and efficiency of researchers working with each database.

2. Background Information

2.1. Context and Landscape

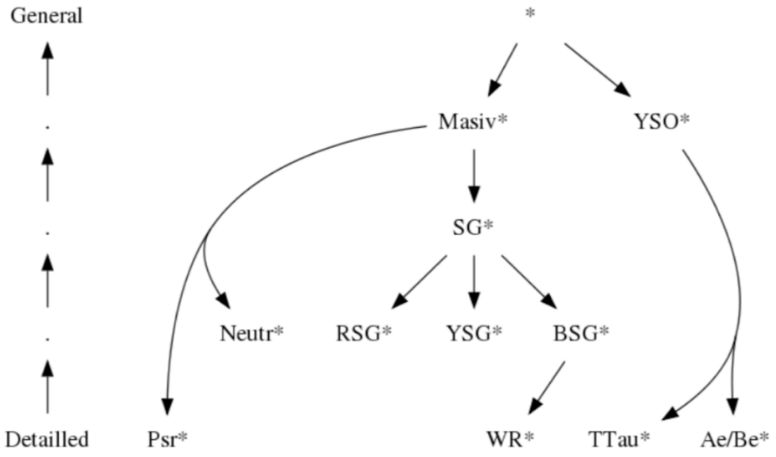
There are many astronomical knowledge and data resources available online to both researchers and the general public. Some of these resources house journal articles, others large compilations of astronomical datasets, which may have specific interests or be very broad in scope. Many of these data have been organized and compiled over the years into various online resources which can be individually searched for the subject or area of interest. However, this has led to many disjoint compilations of research each with their own organizational schemes. The user or researcher is then tasked with searching each one for relevant research information and results which is hopefully relevant to them. This could include digging through papers, re-formatting data, etc. This approach is less than ideal and forces a large amount of work on each user to find the relevant results in a fragmented landscape.

In an effort to organize and unify this information, as well as make it easily accessible across all ranges of astronomical domains, several aggregate databases have been created. The benefits of such compilations allow a researcher to query across a vast array of data from many resources, using search terms which may not have been indexed or catalogued in the original data sets but are now query-able, searchable metadata surrounding the astronomical object(s) of interest. This significantly facilitates the research process, has enabled new kinds of discoveries to be made, and has helped usher in a new era of discovery in astronomy (Mazzarella & the NED Team, 2016).

Two substantive instances of such databases are the "NASA Extragalactic Database" (NED) and the "Set of Identifications, Measurements, and Bibliography for Astronomical Data" (SIMBAD). NED is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration, and—as its name implies—is primarily concerned with organizing information regarding extragalactic objects, or objects outside of our galaxy (About NED, n.d.).

SIMBAD, on the other hand, is managed by the Centre de Données astronomiques de Strasbourg (CDS) and is focused on extrasolar objects (objects outside the solar system) primarily within the Milky way galaxy.

The difference in scope of these two databases is reflected in their respective classification schemes. NED, being broader and more general, is less fine grained in its classification of celestial objects and so it utilizes a simpler 1-dimensional classification system (List of Object Types, n.d.). SIMBAD, having a more sub-galactic "high-resolution" focus, cares a great deal about exactly what an object is and refining that classification as much as possible (Oberto et al., 2018). As a result, it utilizes a hierarchical classification scheme, which looks something like the image below.



Subsection of SIMBAD's Classification Scheme
Categorizations of object types in SIMBAD (Oberto et al., 2018)

2.2. Analysis

A problem that researches often face comes when looking at the same object in each of the two systems: the object may have been assigned different classifications. Since both NED and SIMBAD have pulled their data from a diverse, sprawling array of sources to gather hundreds of millions of objects: cross identification of objects is done between sources in an effort to accurately classify each object. Although extremely advanced and well thought out, this identification process is not perfect. The diversity of the sources being mined for information means significant differences in resources, scope, and concerns between the data. Additionally, any methodology used to classify each object properly is unlikely to have 0% error. Given the sheer scale of such an operation, inevitably there is some degree of discrepancy regarding the classifications given to objects.

This gives rise to doubt and uncertainty in the researcher's mind as they now have to track down the source of the discrepancy to see how best to proceed. For a researcher any degree of uncertainty of the objects class is undesirable as this information's integrity is vital for performing valid inquiry and deriving useful results.

This current pain point is what forms the basis for this proposal. In this paper we propose a potential aid in reducing this level of uncertainty and complexity facing researchers.

3. Proposal Details

3.1. Research Objectives

Research Goal:
Develop an automated system capable of comparing astronomical object classifications assigned by NED and SIMBAD to common objects given by queries of each dataset. The results of the comparison should provide an overview of the level of similarity between the objects compared, the number of possible or candidate matches, and the number of divergent classifications given for each object.

Additionally, each computed comparison done in each object category (i.e. Star), and the classification given by each system (i.e. Blue star) should be displayed to the user.

With this new perspective of a given system's correctness, researchers will have the relevant information needed to further review any differences discovered and proceed accordingly: by either finding the proper classification(s) or moving forward with the knowledge of the classification's accuracy.

Research Objectives:

1. Create intermediate hierarchical classification scheme to relate the linear classification scheme of NED with the hierarchical system of SIMBAD.
2. Encode intermediate hierarchy into concise data scheme which is easily readable and modifiable.

3. Construct a common dictionary where superficial and syntactic differences between SIMBAD and NED will be ignored allowing proper mapping of each entry to its intermediate hierarchy type.
4. Create algorithm to compare new intermediate hierarchical classifications of two common objects from each respective database.
5. Expand systems function from a limited data set and limited type comparisons to any queried dataset for all type comparisons.

3.2. Research Plan

Technical Challenges and Handling Approaches

Relating the classification schemes of NED and SIMBAD will require domain expertise in astronomy to ensure the intermediate hierarchy is complete. The criteria for completeness would be measured by each type of object in both systems—NED and SIMBAD—having a corresponding place in the hierarchy.

Additionally, the level comparisons in the hierarchy must be consistent so as to enable accurate and consistent classification comparisons via parent-child, sibling, or unrelated relationships. This must work for each object class and subclass. SIMBAD for example has 238 unique object classifications, including 16 unique parent classes (Object Classification in SIMBAD, 2013). Our system must work for all of these object types as well as the types present in NED.

To limit the complexity and risk associated with this problem we will first complete only a subset of the hierarchy for a certain object type: stars for example. We will then go through each case of mapping as well as comparing objects from both datasets to ensure each case is handled accurately and gracefully. Once the foundation has been laid for the first object class, further classes will be incorporated into the system.

The second piece to the comparison process will be implementing the utility of actually doing the comparison. Since the algorithm's function will primarily depend on comparing relationships between objects in the intermediate hierarchy, this component of the system also depends heavily on the intermediate hierarchy being complete and meaningful. In general, there will be 3 cases into which comparisons will fall: agreement (same level and same item); basic agreement (common ancestor, or descendant of type); and disagreement (same level but different item, or different tree).

Technology and Implementation Details

Several existing technologies and libraries will be utilized throughout the project. The widely used Astropy Python package and affiliated package Astroquery, will be crucial tools in facilitating the development process. Astropy will be used for astronomy related functions and abstractions and Astroquery for NED and SIMBAD query functionality to automate the retrieval of object information. Astroquery also provides many modules via sub-packages for querying numerous data sources, allowing potential future extendibility (Ginsburg, n.d.)).

Determining whether objects from each database match up with one another will be another function of these packages. We will be defining the same coordinates for both systems, and potentially additional parameters, to ensure that the objects retrieved are a match.

Being that these tools are written in Python, and it is the language of choice for astronomers, it will be our language used for implementation as well (Muna et al., 2016). And of course, NED and SIMBAD will be our modules of choice in comparing classifications.

Potential Threats to Validity

There are some potential threats to the validity of this project and we aim to be scrupulous in addressing all which fall into the scope of our work.

A potential threat to validity could be a frequently occurring “basic agreement” between NED and SIMBAD’s classification systems causing user confusion. Having “basic agreement” means that the objects were not found to be completely inconsistent, but they are not necessarily the same either. In the intermediate hierarchy this relationship would mean sharing a common ancestor. For example, NED may classify an object as a Star and SIMBAD may classify the same object as a horizontal branch star. These classifications are not incompatible, but they are also not enough to surely say that SIMBAD is correct in its labelling. It would seem however to increase the likelihood that the object is in fact a star.

To deal with this potential threat: results should be transparently shown to the user with a clear indication of what information the results convey. The tool will not guarantee an object type is correct; it will instead give you a better idea as to the probability of a given object classification’s correctness.

3.3. Research Methodology

The general philosophy in undertaking the project will be one of iteration and prototyping. We intend to first get basic operational function from our system and build up from there.

As a starting point, some previously developed code will be used and modified as a proof of concept and to further reduce friction of getting progress moving. This code was previously developed by Dr. Pauline Barmby and team, and already facilitates the basic matching of solid or concrete matches by means of literal name comparison (Western Research on Nearby Galaxies - Assembling the big picture, n.d.)).

Instead of first trying to implement full Astroquery integration with user specified object queries, we will instead be using a smaller sample test dataset. We will first be working with a test dataset from M83 Galaxy, which will be obtained by a manual website query of each system. This will be for testing and prototyping purposes.

Additionally, within this sample dataset will we not be looking at all object types; we will start with only a subset of the total number of classifications, then at the end combine all classes to have full functionality. This philosophy will be reflected in the construction of both the comparison algorithm and dictionary which maps each object to the hierarchy; both will start with one object type and build up from there.

Once each object type comparison is working for a given the M83 test dataset, we will then move to comparing objects returned from user specified queries. From here multiple objects including galaxies could be compared. An arbitrary number of objects—including galaxies—could then be selected for comparison.

4. Value of Results & Industry Relevance

For astronomers, knowing the type of an object of interest is of vital importance. If a researcher is trying to uncover a fundamental truth about relationships between celestial objects, how they evolve or form etc. they need to be confident that the information they are using is correct in what it claims to be.

By automating the comparison of object classifications between two of the most significant astronomical databases currently in use, each with their own sophisticated classification methodology, we will be providing that increase in assurance to many potential researchers. They will be notified that either: yes, the data they see is in fact a match; or no, there may be further analysis needed before drawing any conclusions.

More insight from the same information available from these two vast and widely used datasets would increase the efficiency of many researchers' current workflows and provide value to the field as a whole.

5. Bibliography

About NED. (n.d.). Retrieved from NASA/IPAC Extragalactic Database: https://ned.ipac.caltech.edu/Documents?page=Overview

Ginsburg, A. (n.d.). *Accessing Online Astronomical Data*. Retrieved from Astropy: http://www.astropy.org/astroquery/

List of Object Types. (n.d.). Retrieved from NASA/IPAC Extragalactic Database: https://ned.ipac.caltech.edu/?q=help/srcnom/list-objecttypes&popup=1

Mazzarella, J. M., & the NED Team. (2016). Evolution of the NASA/IPAC Extragalactic Database (NED) into a Data Mining Discovery Engine. *Proceedings of the International Astronomical Union*, *12*(S325), 379–384. https://doi.org/10.1017/S1743921316013132

Muna, D., Alexander, M., Allen, A., Ashley, R., Asmus, D., Azzollini, R., … Zonca, A. (2016). The Astropy Problem. *ArXiv:1610.03159 [Astro-Ph, Physics:Physics]*. Retrieved from http://arxiv.org/abs/1610.03159

Oberto, A., Loup, C., Allen, M., Bot, C., Cambrésy, L., Derrière, S., … Vollmer, B. (2018). Categorisations of object types in SIMBAD. *EPJ Web of Conferences*, *186*, 12009. https://doi.org/10.1051/epjconf/201818612009

Object Classification in SIMBAD. (2013, Nov 7). Retrieved from CDS - Strasbourg astronomical Data Center: http://cds.u-strasbg.fr/cgi-bin/Otype?Star

Western Research on Nearby Galaxies - Assembling the big picture. (n.d.). Retrieved from https://nearby-galaxies.github.io