Query Languages for Domain Specific Information from PTF Astronomical Repository

Yilang Wu[1] and Wanming Chu[2]

[1] Computer Networks Laboratory, Aizu University,
Aizu-Wakamatsu, 965-8580, Japan
d8152103@u-aizu.ac.jp

[2] Database Laboratory, Aizu University,
Aizu-Wakamatsu, 965-8580, Japan
w-chu@u-aizu.ac.jp

Abstract. The increasing availability of vast amount of astronomical repositories on the cloud has enhanced the importance of query language for the domain-specific information. The widely used keyword-based search engines (such as Google or Yahoo), fail to suffice for the needs of skilled/semi-skilled users due to irrelevant returns. The domain specific astronomy query tools (such as Astroquery, CDS Portal, or XML) provide a single entry point to search and access multiple astronomical repositories, however these lack easy query composition tools in unit-step or multi-stages query. Based on the previous research studies on domain-specific query language tools, we aim to implement a query language for obtaining the domain-specific information from the astronomical repositories (such as PTF data).

Keywords: Astronomical Information, Domain-specific Information, Multi-stage Query Language.

1 Introduction

Astronomy is now a data-intensive science. Seeking astronomical information through queries is gaining importance in the astronomical domain. The widely used keyword-based search engines such as Google, Yahoo fail to suffice the needs of the astronomy workers (who are well-versed with the domain knowledge required for querying), who have precise queries and expect complete results within time limits (almost real time). And the existing domain-specific searching tools such as Astroquery[1], CDS Portal[2] or XML[3] have not been fully adopted by the current popular astronomical repositories, and access is still based on keyword based search. It is not easy to use. Thus, in this study, we introduce a multi-stage query language for the domain-specific information from the astronomical repositories, such as Palomar Transient Factory (PTF)[4] data.

The proposed multi-stage query language provides a user-level query calculator to formulate a query using domain concepts. These will simplify the querying tasks for the expert and novice domain users. It will enable them to get the desired results[5].

W. Chu et al. (Eds.): DNIS 2015, LNCS 8999, pp. 237–243, 2015.

2 Background

Astronomy is facing a major data avalanche, from multi-terabyte sky surveys and archives, to billions of detected sources, and hundreds of measured attributes per source[6]. The advent of wide-field synoptic imaging has re-invigorated the venerable field of time domain astronomy[7], which involves the study of "how do the astronomical object change with time". It brings new scientific opportunities and also fresh challenges, including handling a huge amount of data storage and transfer, data mining techniques, classification, and heterogeneous data[8][9].

The data overload breeds query tools. The Astroquery[1] is an Astropy (Python Library for Astronomy) affiliated package that contains a collection of tools to access the big tables of online Astronomical data, being advanced in web service specific interfaces. The Strasbourg astronomical Data Center (CDS) is dedicated to the collection and worldwide distribution of astronomical data and related information through HTML, hosting the SIMBAD astronomical database, VizieR catalogue[10] service, and the Aladin[11] interactive software sky atlas[2]. Meanwhile, to seamlessly utilise the highly specialised astronomical datasets or control systems, the XML, for its rich in semantic definition[12], is adopted for a general and highly extensive framework[3].

Our motivation of introducing easy query for the big scientific data in the time domain astronomy was inspired by the iPTF summer school 2014[13], and reports on the Palomar Transient Factory (PTF)[14]. The PTF is a multi-epochal robotic survey of the northern sky that acquires data[4][15] for the scientific study of transient and variable astrophysical phenomena.

3 Astronomical Data Objects for Query

There are variety of astronomical large data repository publicly availabe for download, searching or query, in various formats of tables, image, HTML, XML, or in the form of Map data.

3.1 PTF Data Archive

The PTF data archive is curated by the NASA/IPAC Infrared Science Archive (IRSA). The PTF Level1 data is the initial release (M81, M44, M42, SDSS Stripe 82, and the Kepler Survey Field) in FITS[16] format, including the Epochal (single exposure) Images, and calibrated Photometric Catalogs. The future releases named Level2 data will expand coverage to the entire northern sky, and will include access to the deep coadds (reference images), their associated catalogs, and will include a searchable photometric catalog database. The majority of the released data is at R-band, with a smaller subset at g-band[4].

There are two methods for retrieving PTF data through the IRSA at IPAC: one is an interactive Graphical User Interface (GUI)[17] which is particularly useful as a data exploration tool; and the other is an API (named as IBE) that uses http syntax, providing low-level, program friendly methods for query, including support for the

IVOA Simple Image Access protocol[18]. The query layer of the IRSA Image Server provides the ability to perform spatial and/or relational queries on image metadata tables, with output in IPAC ASCII table, comma-separated-value (CSV), or tab-separated-value (TSV) format. Queries are performed by appending a query string to a base URL identifying the table to query.

3.2 Astroquery

The Astroquery[1] is a set of tools for querying astronomical web forms and databases. All Astroquery modules are supposed to follow the same API, for instance a simplest form of query can be based on coordinates or object names. Most of the modules have been completed using a common API, such as the SIMBAD[19], VizieR[10], IRSA[15] . query modules. The Astroquery serves data as catalogs, archives, simulation data or some other type data, e.g., line list and atomic/molecular cross section and collision rate service. The Astroquery API is as class method in Python. The methods involve query object by name, query region around a coordinate[20].

3.3 SCP Union

The Supernova Cosmology Project (SCP) ”Union2.1”[21][22][23] SN Ia compilation [24] brings together data for 833 SNe, drawn from 19 datasets. Of these, 580 SNe pass usability cuts. All SNe were fit using a single lightcurve fitter (SALT2-1) and uniformly analyzed. The data objects in the SCP Union repository involve figures , cosmology tables, lightcurve data[25]. The figure data illustrates $\Omega_m - \Omega_\wedge$, $\Omega_m - w$, Binned w, Binned ρ obtained from CMB, BAO, and SCP Union2.1 constraints, and also the Binned Hubble Diagram and Residuals. The cosmology tables include the Union Compilation Magnitude vs. Redshift Table, the Covariance Matrix, the Full Table of All SNe, and the CosmoMC Code for Implementing Union Compilation. The Lightcurve Data consists of the SCP HST Cluster Survey Supernova Photometry, the SCP High-z 01 Lightcurve Data and Filters, SNe Summary Table[25].

3.4 XML Based Astronomical Data

The scientific need for a homogenous remote telescope image request system is rapidly escalating as more remote or robotic telescopes are brought to function and scientific programs are created or adapted to use such powerful telescopes. To fit the need, the Remote Telescope Markup Language (RTML)[26] embeds traditional astronomical features such as coordinates and exposure times, and allows for prioritised queue scheduling of telescopes while protecting the telescope operating system. The VOEvent[12] is an international XML standard, defined by the IVOA[27], for transmitting information about a recent astronomical transient, with a view to rapid follow-up, such as Skyalert[28].

3.5 Astronomical Linked Data

Astronomical data artifacts and publications exist in disjointed repositories. The conceptual relationship that links data and publications is rarely made explicit.

Seamless Astronomy Group[29][30][31] at the Harvard-Smithsonian Center for Astrophysics tries to let the connections between literature and data grow more seamless and invisible, so that a researcher can spend more time thinking about science, and less about finding information.

4 Query Language Interfaces

Scientific data within the web data resources is often represented in XML or a related form. XML serves an important role. It has been adopted as the standard language for representing structured data for the traditional Web resources. Thus, many Web-based knowledge management repositories store data and documents in XML. Further, the semantics about the data can be represented by modeling these, with an ontology. Then, it is possible to extract knowledge[32]. Ontologies play an important role in realizing the semantic Web, wherein data will be more sharable because their semantics will be represented in Web-accessible ontologies. Recent reports implement an architecture for this ontology using de facto languages of the semantic Web including OWL and RuleML, thus preparing the ontology for use in data sharing[32]. The users of data resouces are often not skilled in the use of programming languages. These users differ from the Web users and database users. Most of the existing document repositories on the Web have alphabetical and keyword based searches. These are not sufficient for the expert users with precise and complex queries, who require in-depth results within time constraints. Their information needs can be supported by providing user-level schema. Such a schema can support database-style high-level query languages over these repositories. Seeking specialized domain-specific information through queries is gaining importance[5].

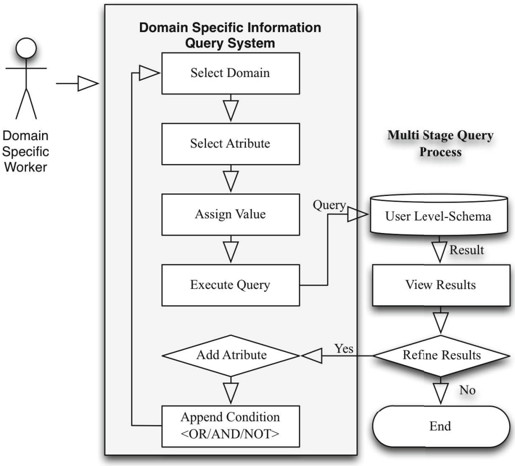

Fig. 1. Query Formation via the Proposed Muti-stage Query Language on User Level

The aim of the is to develop a domain-specific multi-stage query language described in [33] for the archival portals in the scientific domain. Figure 1, represents the query formulation process for astronomical queries, over astronomical objects, through the proposed multi-stage query language. At each stage of query formulation the user can dynamically select an object or a concept to query. Assign a value for it and then either execute the query or further refine the query by adding another attribute(s) and view results. The query is executed on the user-level schema. It provides the users with the segment-level results. Hence, the proposed query language can allow the user to formulate complex DB-style queries using a simple interface and understandable attributes. Figure 1, shows the steps in a typical query formulation process.

5 Summary and Conclusions

A proposed query language for PTF is capable of performing multi-stage visual query (simple, middle, complex, and recursive). We aim to demonstrate a procedure about — how to overcome the existing shortcomings by: (1) implementing the query language to seek diagnostic or hypothesis-directed information (followed by a astronomical domain-expert) and (2) presenting the relevant areas (granular results) of catalog tables or web documents that match the user's query criteria.

In order to achieve the desired query language interfaces for in depth query by skilled and semi-skilled users, it is necessary to organise data as objects in a new schema. Similarly, a set of user level operations, need to be supported. These operations work in a single step at a time. These can be applied in a sequence (by the users) to achieve programmable query language results.

References

1. Ginsburg, A., Robitaille, T., Parikh, M., Deil, C., Mirocha, J., Woillez, J., Svoboda, B., Willett, K., Allen, J.T., Grollier, F., Persson, M.V., Shiga, D.: Astroquery v0.1. 09 (2013)
2. CDS Portal, `http://cdsportal.u-strasbg.fr` (accessed January 2015)
3. Ames, T., Koans, L., Sall, K., Warsaw, C.: Using XML and Java for telescope and instrumentation control. Advanced Telescope and Instrumentation Control Software 4009, 2–12 (2000)
4. Laher, R.R., Surace, J., Grillmair, C.J., Ofek, E.O., Levitan, D., Sesar, B., van Eyken, J.C., Law, N.M., Helou, G., Hamam, N., Masci, F.J., Mattingly, S., Jackson, E., Hacopeans, E., Mi, W., Groom, S., Teplitz, H., Desai, V., Hale, D., Smith, R., Walters, R., Quimby, R., Kasliwal, M., Horesh, A., Bellm, E., Barlow, T., Waszczak, A., Prince, T., Kulkarni, S.R.: IPAC Image Processing and Data Archiving for the Palomar Transient Factory (April 2014)
5. Madaan, A., Chu, W.: Handling domain specific document repositories for application of query languages. In: Madaan, A., Kikuchi, S., Bhalla, S. (eds.) DNIS 2014. LNCS, vol. 8381, pp. 152–167. Springer, Heidelberg (2014), `http://dx.doi.org/10.1007/978-3-319-05693-7_10`
6. Djorgovski, S.G.: New Astronomy With a Virtual Observatory Astronomy is Facing a Major

7. Kasliwal, M.: Transients in the Local Universe: Today and Tomorrow. In: International Conference on Computational Physics, Singapore (2015)
8. Graham, M.J., Djorgovski, S.G., Mahabal, A., Donalek, C., Drake, A., Longo, G.: Data challenges of time domain astronomy. Distributed and Parallel Databases 30, 371–384 (2012)
9. Madaan, A., Chu, W., Bhalla, S.: VisHue: Web page segmentation for an improved query interface for medlinePlus medical encyclopedia. In: Kikuchi, S., Madaan, A., Sachdeva, S., Bhalla, S. (eds.) DNIS 2011. LNCS, vol. 7108, pp. 89–108. Springer, Heidelberg (2011), `http://dx.doi.org/10.1007/978-3-642-25731-5_9`
10. Ochsenbein, F., Bauer, P., Marcout, J.: The VizieR database of Astronomical Catalogues. 10, 10 (2000)
11. Franke, B.: An Introduction To The Aladin Sky Atlas. Tech. rep.
12. Seaman, R., Williams, R., Optical, N., Observatory, A., Allan, A., Barthelmy, S., Bloom, J.S., Brewer, J.M., Denny, R.B., Fitzpatrick, M., Graham, M., Gray, N., Hessman, F., Marka, S., Rots, A., Vestrand, T., Wozniak, P.: VOEvent reporting metadata. pp. 1–27 (2011)
13. iptf summer school, `http://phares.caltech.edu/iptf/iptf_SummerSchool_2014/` (accessed august 2014)
14. Law, N.M., Kulkarni, S.R., Dekany, R.G., Ofek, E.O., Quimby, R.M., Nugent, P.E., Surace, J., Grillmair, C.C., Bloom, J.S., Kasliwal, M.M., Bildsten, L., Brown, T., Cenko, S.B., Ciardi, D., Croner, E., Djorgovski, S.G., van Eyken, J.C.: a. V. Filippenko, D. B. Fox, a. Gal-Yam, D. Hale, N. Hamam, G. Helou, J. R. Henning, D. a. Howell, J. Jacobsen, R. Laher, S. Mattingly, D. McKenna, a. Pickles, D. Poznanski, G. Rahmer, a. Rau, W. Rosing, M. Shara, R. Smith, D. Starr, M. Sullivan, V. Velur, R. S. Walters, and J. Zolkower, "The Palomar Transient Factory: System Overview, Performance and First Results," p. 12 (2009)
15. "IRSA Image Server", `http://irsa.ipac.caltech.edu/ibe/index.html` (accessed: January 2015)
16. Pence, W.D., Chiapetti, L., Page, C.G., Shaw, R.A., Stobie, E.: Definition of the Flexible Image Transport System (FITS), Version 3.0. vol. 42 (2010)
17. "PTF GUI", `http://irsa.ipac.caltech.edu/applications/ptf/` (accessed: January 2015)
18. Access, S.I., Tody, D., Dowler, P., Dowler, P., Tody, D., Plante, R.: IVOA Simple Image Access IVOA Proposed Recommendation 2014-07-07, 1–25 (2014)
19. SIMBAD: Set of Identifications, Measurements and Bibliography for Astronomical, `http://simbad.u-strasbg.fr/Simbad` (accessed: January 2015)
20. Astroquery, `https://astroquery.readthedocs.org` (accessed: January 2015)
21. Suzuki, N., Rubin, D., Lidman, C., Aldering, G., Amanullah, R., Barbary, K., Barrientos, L.F., Botyanszki, J., Brodwin, M., Connolly, N., Dawson, K.S., Dey, A., Doi, M., Donahue, M., Deustua, S., Eisenhardt, P., Ellingson, E., Faccioli, L., Fadeyev, V., Fakhouri, H.K., Fruchter, A.S., Gilbank, D.G., Gladders, M.D., Goldhaber, G., Gonzalez, A.H., Goobar, A., Gude, A., Hattori, T., Hoekstra, H., Hsiao, E., Huang, X., Ihara, Y., Jee, M.J., Johnston, D., Kashikawa, N., Koester, B., Konishi, K., Kowalski, M., Linder, E.V., Lubin, L., Melbourne, J., Meyers, J., Morokuma, T., Munshi, F., Mullis, C., Oda, T., Panagia, N., Perlmutter, S., Postman, M., Pritchard, T., Rhodes, J., Ripoche, P., Rosati, P., Schlegel, D.J., Spadafora, A., Stanford, S.A., Stanishev, V., Stern, D., Strovink, M., Takanashi, N., Tokita, K., Wagner, M., Wang, L., Yasuda, N., Yee, H.K.C.: The Hubble Space Telescope Cluster Supernova Survey: V. Improving the Dark Energy Constraints Above z¿1 and Building an Early-Type-Hosted Supernova Sample 27 (2011)

22. Amanullah, R., Lidman, C., Rubin, D., Aldering, G., Astier, P., Barbary, K., Burns, M.S., Conley, A., Dawson, K.S., Deustua, S.E., Doi, M., Fabbro, S., Faccioli, L., Fakhouri, H.K., Folatelli, G., Fruchter, A.S., Furusawa, H., Garavini, G., Goldhaber, G., Goobar, A., Groom, D.E., Hook, I., Howell, D.A., Kashikawa, N., Kim, A.G., Knop, R.A., Kowalski, M., Linder, E., Meyers, J., Morokuma, T., Nobili, S., Nordin, J., Nugent, P.E., Ostman, L., Pain, R., Panagia, N., Perlmutter, S., Raux, J., Ruiz-Lapuente, P., Spadafora, A.L., Strovink, M., Suzuki, N., Wang, L., Wood-Vasey, W.M., Yasuda, N.: Spectra and Light Curves of Six Type Ia Supernovae at 0.511 ¡ z ¡ 1.12 and the Union2 Compilation, 33 (2010)
23. Kowalski, M., Rubin, D., Aldering, G., Agostinho, R.J., Amadon, A., Amanullah, R., Balland, C., Barbary, K., Blanc, G., Challis, P.J., Conley, A., Connolly, N.V., Covarrubias, R., Dawson, K.S., Deustua, S.E., Ellis, R., Fabbro, S., Fadeyev, V., Fan, X., Farris, B., Folatelli, G., Frye, B.L., Garavini, G., Gates, E.L., Germany, L., Goldhaber, G., Goldman, B., Goobar, A., Groom, D.E., Haissinski, J., Hardin, D., Hook, I., Kent, S., Kim, A.G., Knop, R.A., Lidman, C., Linder, E.V., Mendez, J., Meyers, J., Miller, G.J., Moniez, M., Mourao, A.M., Newberg, H., Nobili, S., Nugent, P.E., Pain, R., Perdereau, O., Perlmutter, S., Phillips, M.M., Prasad, V., Quimby, R., Regnault, N., Rich, J., Rubenstein, E.P., Ruiz-Lapuente, P., Santos, F.D., Schaefer, B.E., Schommer, R.A., Smith, R.C., Soderberg, A.M., Spadafora, A.L., Strolger, L.G., Strovink, M., Suntzeff, N.B., Suzuki, N., Thomas, R.C., Walton, N.A., Wang, L., Wood-Vasey, W.M., Yun, J.L.: Improved Cosmological Constraints from New, Old and Combined Supernova Datasets 49 (2008)
24. Supernova Cosmology Project Union 2.1 Compilation, `http://supernova.lbl.gov/union/` (accessed: January 2015)
25. Data Description for Supernova Cosmology Project, `http://supernova.lbl.gov/union/descriptions.html` (accessed: January 2015)
26. Pennypacker, C., Boer, M., Denny, R., Hessman, F.V., Aymon, J., Duric, N., Gordon, S., Barnaby, D., Spear, G., Hoette, V.: RTML - a standard for use of remote telescopes. Astronomy and Astrophysics 395(2), 727–731 (2002)
27. Documents and Standards about Virtual Observatory, `http://www.ivoa.net/documents/index.html` (accessed: January 2015)
28. Williams, R.D., Djorgovski, S.G., Drake, A.J., Graham, M.J., Mahabal, A.: Skyalert: Real-time Astronomy for You and Your Robots XXX, 4 (2009)
29. Alyssa Goodmans, Seamless Astronomy, tech. rep. (2013)
30. Seamless Astronomy, `http://projects.iq.harvard.edu/seamlessastronomy` (accessed: January 2015)
31. Goodman, A., Fay, J., Muench, A., Pepe, A., Udomprasert, P., Wong, C.: World-Wide Telescope in Research and Education. eprint arXiv:1201.1285, 4 (2012)
32. Kim, H., Sengupta, A.: Extracting knowledge from xml document repository: a semantic web-based approach. Information Technology and Management 8(3), 205–221 (2007)
33. Madaan, A., Bhalla, S.: Domain specific multistage query language for medical document repositories. Proc. VLDB Endow. 6(12), 1410–1415 (2013)