

Detección de cúmulos estelares en galaxias cercanas utilizando técnicas de Machine Learning y algoritmos de aplicación en redes sociales

Universidad Tecnológica Nacional
Facultad Regional Córdoba

Informe de avance - **AÑO 2023**

Tesista: Esp. Ing. Martin Casatti

Director: Dr. Marcelo Marciszack

El presente es el informe de avance de la Tesis de Doctorado en Ingeniería, mención Sistemas de Información, titulada “Detección de cúmulos estelares en galaxias cercanas utilizando técnicas de Machine Learning y algoritmos de aplicación en redes sociales”, cuyo plan de Tesis ha sido aprobado según resolución N° 606/2023, de fecha 26 de Abril de 2023, siendo los directores de la misma el Dr. Ing. Marcelo Marciszack (DIRECTOR), y el Dr. Carlos Feinstein (CO-DIRECTOR).

Índice general

Índice general	2
1. Cursos y seminarios realizados	2
2. Avances teóricos y metodológicos realizados	2
3. Participación en eventos académicos	3
4. Otras actividades que considere importante consignar	4
5. Dificultades durante el presente año de trabajo	5
6. Cronograma del Plan de Trabajo para 2024	5
Referencias	5

1. Cursos y seminarios realizados

Durante el presente año no se han realizado cursos y capacitaciones de posgrado con validez para esta Tesis.

2. Avances teóricos y metodológicos realizados

Relevamiento de fuentes de datos

Se realizó un relevamiento detallado de las posibles fuentes de datos astronómicos y de redes sociales para su utilización durante el desarrollo de la tesis.

En dicho análisis preliminar se tuvieron en cuenta criterios tales como la accesibilidad de la información, la cantidad de datos, el tipo de atributos de los datos almacenados etc. Dentro de los más prometedores se encontraron los siguientes:

VVV: Un proyecto de la Organización Europea para la Investigación Astronómica del hemisferio Sur, VVV¹ funciona desde 2010 y ya lleva detectados más de 350 cúmulos estelares[3].

LSST: Generado por el observatorio Vera C. Rubin, emplazado en Chile, LSST² generará aproximadamente 20 terabytes de información por noche, durante los 10 años de duración prevista para el proyecto[4, 5].

¹<https://vvvsurvey.org/>

²<https://www.lsst.org/>

VISCACHA: El proyecto VISCACHA es un estudio fotométrico diseñado específicamente para cúmulos estelares en la Pequeña y Gran Nube de Magallanes[6].

Con respecto a las fuentes de datos sobre redes sociales, se accedió a Network Repository³ que es un repositorio curado de diferentes datos en formato de grafo, destinado a tareas de investigación y testing de algoritmos[2], así como al Stanford Large Network Dataset Collection⁴, con similares características y mantenido por la universidad de Stanford.

Dentro del mismo se encuentran varios sets de datos de redes sociales, de los cuales los más prometedores resultaron los siguientes:

fb-pages-artists: Base de datos con relaciones recíprocas entre artistas⁵. Cuenta con más de 50.000 nodos y 800.000 relaciones.

soc-linkedin: Base de datos con información de LinkedIn⁶.

com-Youtube: Base de datos para el estudio de comunidades en Youtube⁷. Cuenta con aproximadamente 1 millón de nodos y casi 3 millones de relaciones[7].

Análisis y selección de base de datos de soporte

Ante la compra de la base de datos propuesta (OrientDB) por parte de la firma SAP y el congelamiento del proyecto, se resolvió reemplazar el almacenamiento previsto por un proyecto que se encuentre activo, en disponibilidad y con posibilidades de desarrollo durante el tiempo de duración de la tesis.

A tal efecto se evaluó el uso de ArcadeDB, el que tiene como origen una derivación directa de OrientDB, el mismo conjunto de funcionalidad y una base de código activa, en desarrollo y modernizada en comparación al proyecto original[8].

Algoritmos de detección de clusters, estudio sistemático

Se comenzó la elaboración de un estudio sistemático de los algoritmos actualmente existentes para la detección de clusters y agrupaciones, tanto en el ámbito astronómico[1] como de redes sociales y de propósito general[9].

3. Participación en eventos académicos

CIACA2023

Con fecha 6/9/2023 se recibió la notificación de la aceptación del paper “Detección de cúmulos estelares en galaxias cercanas utilizando técnicas de Machine Learning y

³<https://networkrepository.com/>

⁴<https://snap.stanford.edu/data/>

⁵<https://networkrepository.com/fb-pages-artist.php>

⁶<https://networkrepository.com/soc-linkedin.php>

⁷<https://snap.stanford.edu/data/com-Youtube.html>

algoritmos de aplicación en redes sociales’’, dentro de la categoría Reflection Paper, para su publicación en la 10ª Conferencia Ibero Americana de Computación Aplicada (CIACA2023)⁸, desarrollada durante los días 22 y 23 de Octubre de 2023, en Madeira Portugal.

CoNaIISI 2023

Se aprobó la publicación de un paper titulado ‘‘Indicadores cuantitativos preliminares sobre la investigación estudiantil en CoNaIISI’’, en el congreso CoNaIISI 2023⁹. El tesista participa en calidad de autor y es el responsable, en el proyecto de investigación, de la implementación de la base de datos de grafos que da soporte a las estadísticas e indicadores, siendo ésta la misma tecnología que se utilizará en el desarrollo de la tesis para la implementación del almacenamiento de datos astronómicos bajo análisis.

4. Otras actividades que considere importante consignar

X Congreso Nacional de Extensión Universitaria

Realizado durante los días 29, 30 y 31 de Marzo de 2023, en la Universidad Nacional de La Pampa¹⁰, el tesista participó en charlas y conversatorios, como miembro de Room 101, el Grupo de Estudios sobre Ciencia y Ficción dependiente de la Secretaría de Extensión Universitaria de UTN Facultad Regional Córdoba, exponiendo sobre técnicas no tradicionales de divulgación científica y tecnológica y aplicaciones del género de Ciencia Ficción como herramienta didáctica en educación superior.

CIRE 2023

El tesista dictó un taller, durante la III edición del Congreso Internacional de Robótica Educativa¹¹, sobre los usos de la ciencia ficción como generadora de ideas y problemas que luego se pueden volcar en el diseño e implementación de robots. La jornada tuvo lugar los días 1 y 2 de Junio de 2023, en la Universidad Tecnológica Nacional, Facultad Regional Córdoba.

Pórtico 7.23 - Universidad Nacional de La Plata

Durante los días 10 y 11 de Noviembre de 2023, el tesista dictó un taller sobre Inteligencia Artificial y Ética (10 de noviembre) y una Mesa de Debate sobre Inteligencia Artificial y su uso en educación (11 de noviembre), la cual coordinó y moderó. Todo eso en el marco del Encuentro Pórtico 7.23, desarrollado en la Facultad de Ingeniería de la Universidad Nacional de La Plata.

⁸<https://ciaca-conf.org/>

⁹<https://frtutn.cloud/conaiisi/>

¹⁰<https://www.unlpam.edu.ar/XCongresoExtension/>

¹¹<https://educacion.cordoba.gob.ar/congreso-internacional-de-robotica-educativa/>

5. Dificultades durante el presente año de trabajo

No se presentaron dificultades particulares para el desarrollo de las tareas previstas durante el presente año.

6. Cronograma del Plan de Trabajo para 2024

Para el período 2024 se propone el siguiente plan de trabajo (Figura 1):

1. Mapeo sistemático de literatura relacionada a algoritmos de clustering
2. Construcción de infraestructura de almacenamiento y carga de datos de prueba
3. Aplicación de algoritmos de clustering astronómico y comparación con clusters conocidos
4. Estudio de atributos en grafos de redes sociales para su mapeo como atributos astronómicos
5. Publicación de resultados obtenidos

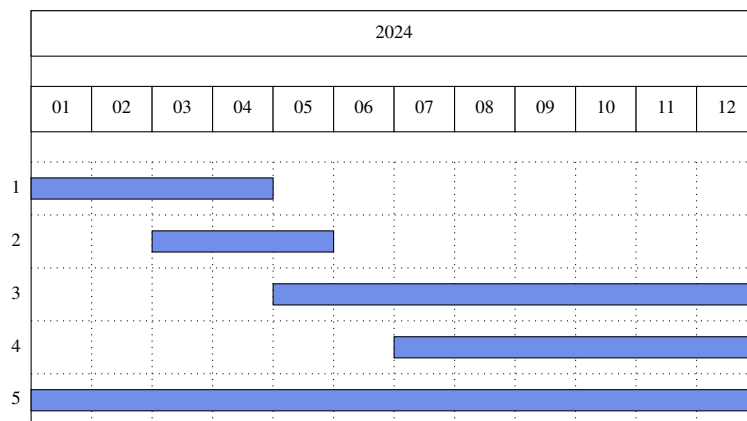
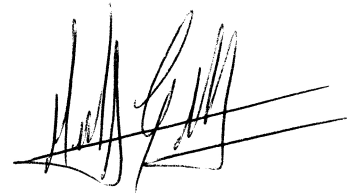


Figura 1: Año 2024

Referencias

1. Schmeja, S. Identifying star clusters in a field: A comparison of different algorithms. *Astronomische Nachrichten* **332**, 172-184 (2011).
2. Rossi, R. A. y Ahmed, N. K. *The Network Data Repository with Interactive Graph Analytics and Visualization* en AAAI (2015). <https://networkrepository.com>.
3. Borissova, J. *et al.* New Galactic star clusters discovered in the VVV survey. *Astronomy & Astrophysics* **532**, A131 (2011).

4. Tyson, J. A. Large synoptic survey telescope: overview. *Survey and Other Telescope Technologies and Discoveries* **4836**, 10-20 (2002).
5. Jurić, M. *et al.* The LSST data management system. *arXiv preprint arXiv:1512.07914* (2015).
6. Maia, F. F. *et al.* The VISCACHA survey–I. Overview and first results. *Monthly Notices of the Royal Astronomical Society* **484**, 5702-5722 (2019).
7. Yang, J. y Leskovec, J. *Defining and Evaluating Network Communities based on Ground-truth* 2012. arXiv: [1205.6233 \[cs.SI\]](https://arxiv.org/abs/1205.6233).
8. *ArcadeDB Manual* [Online; accessed 18. Nov. 2023]. <https://docs.arcadedb.com>.
9. Ahmad, A. y Khan, S. S. Survey of state-of-the-art mixed data clustering algorithms. *Ieee Access* **7**, 31883-31902 (2019).



Ing. Martin Gustavo Casatti