Bases de datos de grafos como soporte para la detección de estrellas jóvenes en cúmulos estelares cercanos

Martín Casatti[1], Analía Guzmán[1], María Alejandra Paz Menvielle[1]

Universidad Tecnológica Nacional - Facultad Regional Córdoba, Córdoba, Argentina,
mcasatti@frc.utn.edu.ar, aguzman@frc.utn.edu.ar,pazmalejandra@gmail.com

Resumen El presente trabajo explorará las características más destacadas que debe reunir una base de datos de grafos para la implementación de un sistema de reconocimiento de patrones estructurales en cúmulos estelares cercanos. Se analizarán los parámetros a almacenar, de acuerdo a la información relevada por diversos proyectos de observación astronómica, así como el modelo de representación más prometedor, considerando la finalidad del almacenamiento. Se describirán también algunas de las características de bases de datos existentes, así como sus ventajas y desventajas a la hora de implementar el mencionado sistema de reconocimiento. Se demostrará que una representación en forma de grafos de la información existente en los repositorios astronómicos no sólo es posible y puede ayudar al procesamiento automático de dicha información sino que además puede proveer un mecanismo efectivo para la implementación de algoritmos de reconocimiento de patrones y la consecuente detección de estructuras estelares de interés.

Keywords: grafos, base de datos, modelado, detección de patrones, astronomía, cúmulos estelares

1. Contexto

El presente informe forma parte de los trabajos orientados a la elaboración de una tesis de maestría para la obtención del título de Magister en Ingeniería en Sistemas de Información, la cual busca determinar la efectividad de algoritmos basados en grafos para la detección de patrones estructurales en modelos de sistemas astronómicos, específicamente en lo que respecta a cúmulos estelares en galaxias cercanas.

La mencionada tesis se desarrolla en el marco de un proyecto de investigación y desarrollo que ha sido homologado por la Secretaría de Investigación, Desarrollo y Posgrado de la Universidad Tecnológica Nacional, el cual se lleva a cabo en el ámbito del CIDS – Centro de Investigación, Desarrollo y Transferencia en Sistemas de Información.

En la actualidad existe una gran cantidad de información de las galaxias cercanas debido, en gran parte, a que el Telescopio Espacial Hubble (HST) ha permitido obtener datos con alta resolución espacial utilizando varias cámaras de

campo amplio (WFPC2, ACS)[12]. Este acceso facilita realizar investigaciones vinculadas con agrupaciones estelares, sobre diferentes poblaciones e historias de formación estelar.

Desde hace tiempo se reconoce en el campo de la astrofísica que los cúmulos estelares son laboratorios importantes para la investigación, ya que contienen muestras estadísticamente significativas de estrellas de aproximadamente la misma edad en un espacio reducido. Por otra parte, las agrupaciones estelares existentes en los mismos brindan información valiosa para la comprensión de la estructura de la galaxia que las contiene.

Existe actualmente una enorme cantidad de información obtenida a partir de proyectos de observación continua, principalmente en modo "survey"[6, 15] como pueden ser:

- VVV[24] (https://vvvsurvey.org/)
- LSST[19] (https://www.lsst.org/)
- SDSS[7] (https://www.sdss.org/)
- Gaia-ESO[17] (https://www.gaia-eso.eu/)

Todos los ejemplos mencionados requieren de mecanismos automáticos para el análisis de los datos.

Los algoritmos de "Data Mining" (DM), en particular los relacionados con el reconocimiento automático de patrones, están en la actualidad teniendo una importante revisión y desarrollo[5, 2, 28] para su aplicación sobre los datos que surgen de los grandes "surveys".

En este trabajo se pretende exponer las capacidades de las bases de datos de grafos para soportar algoritmos de reconocimiento automático de patrones sobre datos astronómicos. Se busca determinar si dichas técnicas, provenientes de otros ámbitos de aplicación, son trasladables al ámbito de la astronomía, si se requieren o no adecuaciones, o si la información de las bases de datos astronómicos requiere de algoritmos de reconocimiento diseñados específicamente para las mismas.

En una primera etapa se analizará la forma en que se realizará el preprocesamiento y adquisición de datos, para luego realizar las etapas de extracción de las características fundamentales y el agrupamiento o clasificación para lograr la identificación y parametrización de nuevas agrupaciones estelares.

Respecto al preprocesamiento y adquisición de datos se describirá como es el modelo de datos seleccionado y como a partir de la muestra de datos obtenida del instrumental óptico, se lo preparara para almacenarlo en un grafo, para luego poder continuar con las etapas de detección y reconocimiento de patrones relacionados.

2. Introducción

Un patrón es una entidad a la que se le puede dar un nombre y que está representada por un conjunto de propiedades medidas y las relaciones entre ellas representadas en el denominado *vector de características*[32].

En el dominio de los datos astronómicos un patrón puede ser las distancias medias entre estrellas del mismo cúmulo, sus características espectrales, su curva de luminosidad, etc. El vector de características estaría conformado, en este caso, por características físicas, químicas o estructurales que relaciones los distintos elementos.

Existen trabajos que se han focalizado en mejorar el conocimiento de nuestra propia Galaxia y de las Nubes de Magallanes, por ejemplo Baume et al. 2008[3], pero actualmente hay varios factores que incrementan de forma importante tanto la cantidad de objetos a analizar tornando especialmente relevante un método automático para el procesamiento de los datos.

El reconocimiento automático, descripción, clasificación y agrupamiento de patrones son actividades importantes en una gran variedad de disciplinas científicas, como biología, psicología, medicina, visión por computador, inteligencia artificial, teledetección, etc. La principal importancia que tiene la detección de patrones en los datos es que se pueden inferir causas para la agrupación de los mismos y es aquí donde radica el interés de la aplicación de éstas técnicas al ámbito astronómico.

Si bien actualmente se están investigando métodos de reconocimiento basados en máquinas de soporte vectorial[9] el enfoque utilizado en la presente línea de investigación se centra en las características estructurales de los grafos[8, 25].

Se describen a continuación conceptos fundamentales relacionados a la construcción de grafos:

Nodo: Un nodo es un *entidad de información* diferente de cualquier otra entidad en el modelo. Todos los nodos representan un único conjunto de información la cual es indivisible e independiente de cualquier otra información representada. Unidades mayores pueden representarse únicamente como grupos de nodos relacionados.Para los fines que se persiguen es necesario que cada nodo pueda ser identificado de manera unívoca y diferente de todos los demás nodos de la red.

Enlace: Un enlace se utiliza para establecer una relación entre dos nodos del modelo. Un enlace solo puede conectar dos nodos. Uno denominado *origen* desde el que sale el enlace y uno denominado *destino* al cual llega. Una relación entre un nodo de origen y dos nodos de destino requiere de dos enlaces, ambos con el mismo origen, pero con diferentes destinos. Un nodo puede ser origen de varios enlaces, así como puede ser destino de varios enlaces[31, 4].

2.1. Patrones en Grafos

El reconocimiento o detección de patrones dentro de grafos busca detectar un subgrafo (patrón) en un grafo (objetivo). Debemos considerar que esta búsqueda de coincidencias se puede descomponer en dos partes:

1. Una concordancia estructural, en donde los nodos y relaciones del patrón conforman una estructura existente en el grafo objetivo.
2. Una concordancia a nivel de elementos, en donde los nodos y relaciones, a nivel de sus atributos particulares, tiene los mismos valores que en la estructura encontrada en el grafo objetivo.

Muchas veces la búsqueda de estas dos concordancias se ejecuta de forma separada para optimizar los algoritmos o reducir el espacio de búsqueda[14].

En el dominio bajo estudio la detección de un subgrafo (patrón) se realizará sobre los grafos generados a partir de información astronómica suministrada por archivos de observación continua (survey) tales como los encontrados en los proyectos VVV Survey, Large Scale Telescope (LST) o Hubble Space Telescope (HST), siendo éste último el origen de los datos que se van a utilizar en éste y sucesivos trabajos.

2.2. Métricas en grafos

Una herramienta ampliamente utilizada para describir grafos y que muchas veces se utiliza para iniciar el análisis de patrones existentes en los mismos, es el cálculo de métricas[31], locales o globales, que permiten caracterizar el grafo objetivo o el grafo patrón. Las métricas se pueden dividir en dos grandes grupos:

- Métricas estáticas: Cuando se calculan sobre un grafo estático en un punto en el tiempo determinado. Se enfocan principalmente en las características estructurales del mismo.
- Métricas dinámicas: Tienen en cuenta la dimensión temporal de los cambios que se producen sobre el grafo. Están más enfocadas en las variaciones entre dos instantes de tiempo, antes que en las características propias del grafo en cada uno de esos instantes.

Otro enfoque para el análisis de las métricas radica en analizar sobre qué componentes del grafo se realizan las mediciones. Desde este punto de vista se tienen diversas perspectivas, siendo las más comunes:

- Métricas de redes (o globales): Son las métricas que toman como referencia el grafo completo, con todos los nodos y arcos que lo conforman.
- Métricas nodos (o locales): Son aquellas que toman como referencia un nodo o subconjunto de nodos para realizar los cálculos.

A continuación, se detallan algunas métricas más comunes:

Métricas globales: Centralidad: Esta métrica trata de determinar que nodo o nodos ocupan una ubicación central en la red, estando equidistante de los demás nodos.

Conexionado: Busca establecer el grado en el que los nodos de un grafo están conectados con todos los demás nodos del mismo. Se puede encontrar, aplicando esta métrica, componentes fuertemente conectados o débilmente conectados.

Cantidad de Componentes: En un grafo que no es completamente conexto, indica la cantidad de subgrafos conexos que forman parte del grafo. Un componente es un conjunto de nodos conectados que forman parte del grafo principal.

Tamaño del componente gigante: Mide la cantidad de nodos que tiene el componente conectado que es mayor que todos los demás componentes del grafo.

En un grafo conexo el tamaño del componente gigante es igual a la cantidad total de nodos.

Ruta más corta/larga: Expresa la longitud (en arcos) mínima/máxima entre dos nodos dados.

Métricas locales: Conectividad: Expresa la cantidad de conexiones que posee un nodo determinado. Se puede expresar como 'grado', si no tiene en cuenta la dirección de los arcos que inciden o salen del nodo, o como 'grado de entrada' o 'grado de salida' cuando solamente tiene en cuenta los arcos entrantes o salientes, respectivamente.

Centralidad: Es una métrica, asociada a un nodo en un grafo, que determina su importancia relativa dentro de éste, pudiendo dividirse en:

- Centralidad de grado: Cantidad de conexiones con otros nodos
- Centralidad de cercanía: Indica qué tan cerca se encuentra una unidad de la red de otras.
- Centralidad de intermediación: Indica si una unidad se encuentra dentro de algunas de las rutas más cortas que existen entre dos nodos de la red.

2.3. Diseño de reconocimiento de patrones

El objetivo principal de un sistema de reconocimiento automático de patrones es descubrir la naturaleza subyacente de un fenómeno u objeto, describiendo y seleccionado las características fundamentales que permitan clasificarlos en una categoría determinada[29][16].

Sistemas automáticos de reconocimiento de patrones permiten abordar problemas en informática, en ingeniería y en otras disciplinas científicas[13][23], por lo tanto el diseño de cada etapa requiere de criterios de análisis conjuntos para validar los resultados[21][20].

Luego de analizar diferentes formas de diseñar un sistema de reconocimiento de patrones, se consideran tres fases[1]:

1. Adquisición y preproceso de datos.
2. Extracción de características.
3. Toma de decisiones o agrupamiento.

En la fase de Adquisición y preproceso de datos, se preparará la infraestructura de la base de datos para poder continuar con las siguientes fases.

Para las dos siguientes fases, extracción de características y toma de decisiones o agrupamiento se considerarán los parámetros más relevantes que conforman agrupaciones estelares de interés, trabajando en colaboración con expertos del Instituto Astrofísico de La Plata, Buenos Aires, Argentina.

3. Métodos y resultados

El auge de las actuales bases de datos de grafos nos brinda posibilidades importantes a la hora de modelar un dominio para el cual las bases de datos relacionales no tienen aplicación directa.

La posibilidad de incluir en el esquema diseñado, información de tipos variados, sin penalizar por ellos las capacidades de búsqueda o la representatividad de la información es una de las características más favorables del modelo basado en grafos, lo que la hace especialmente recomendable en entornos en donde no se cuenta con un esquema establecido o no es estable o sufre fercuentes variaciones, ejemplos de esto son prototipos de diseño, bases de datos para la prueba de conceptos y almacenamientos para minería de datos[26].

En términos de expresividad las bases de datos reducen la diferencia de impedancia entre el modelo de análisis y la implementación final, un problema que ha acosado a los diversos modelos de bases de datos desde hace muchos años.

3.1. Factores de representación

Para tener información relevante del dominio elegido y poder detectar patrones, se necesita contar con una base de grafos que represente la información de los cumulos estelares con la mayor exactitud posible.

Actualmente se deben tener en cuenta al menos cuatro factores fundamentales a la hora de diseñar un sistema de representación del conocimiento en cualquier dominio dado[30]:

- **Adecuación Representacional:** Habilidad para representar todas las clases de conocimiento que son necesarias en el dominio.
- **Adecuación Inferencial:** Habilidad de manipular estructuras de representación de tal manera que devengan o generen nuevas estructuras que correspondan a nuevos conocimientos inferidos de los anteriores.
- **Eficiencia Inferencial:** Capacidad del sistema para incorporar información adicional a la estructura de representación, llamada metaconocimiento, que puede emplearse para focalizar la atención de los mecanismos de inferencia con el fin de optimizar los cómputos.
- **Eficiencia en la Adquisición:** Capacidad de incorporar fácilmente nueva información. Idealmente el sistema por sí mismo deberá ser capaz de controlar la adquisición de nueva información y su posterior representación.

Estos factores se tuvieron en cuenta durante el proceso de diseño de las estructuras de la base de datos, de manera tal que se puedan maximizar los resultados a la vez que se mantienen eficientes las operaciones de cómputo.

3.2. El modelo de grafos etiquetados

El enfoque propuesto en el presente trabajo se basa en el modelo de grafos etiquetados, existente en multitud de productos de base de datos de grafos, tanto comerciales como Open Source.

Un *grafo de propiedades etiquetadas* esta compuesto de nodos, relaciones, propiedades y etiquetas, tal como se puede apreciar en la Figura 1.

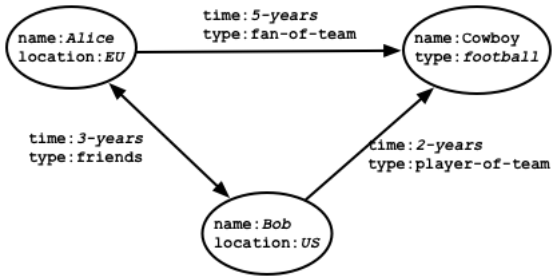

Figura 1. Modelo de Grafo Etiquetado

- Los **nodos** contienen **propiedades**. Se puede pensar en los nodos como documentos que agrupan un conjunto de propiedades que definen las características del documento al que pertenecen.
- Los **nodos** pueden ser marcados con una o más **etiquetas**. Las etiquetas agrupan nodos por características similares o indican los roles que cada uno juega en el modelo de datos.
- Las **relaciones** conectan nodos entre si y definen la estructura de un grafo. Una relación tiene una dirección, un nodo de origen y un nodo de destino.
- Como los nodos las relaciones también pueden tener propiedades que le agregan valor semántico a la relación.
- En todos los casos, las propiedades y sus valores pueden utilizarse para restringir los resultados de las búsquedas realizadas en un grafo.

En base al modelo teórico de grafos etiquetados expuesto anteriormente se diseñó un modelo de implementación que brindará soporte a los datos sobre los que actuarán los algoritmos de reconocimiento.

El elemento central a considerar en el modelo de implementación son los **nodos**. Los nodos, en este ámbito modelan lecturas puntuales obtenidas por el instrumento de observación. Cada nodo se hace equivalente, con este enfoque, a un registro de los archivos de intercambio de información. Este proceso se detallará en la sección 3.3.

Se generará un identificador único para cada nodo para simplificar las consultas y posterior recuperación de la información asociada a los mismos. Existen algunas propiedades inherentes a cada una de las observaciones que serán almacenadas en los nodos correspondientes, como por ejemplo las coordenadas de dicha observación, la sección del cielo que se está relevando y los ajustes del instrumento de observación. Todas estas características son necesarias para reconstruir las condiciones originales de observación a partir de los datos.

Las **etiquetas**, una vez importados los datos generales de cada nodo, pasarán a cumplir un rol fundamental ya que son las responsables del almacenamiento de toda la información adicional que se utilizará en las búsquedas.

Dentro de los posibles valores a almacenar se encuentran las diversas lecturas tomadas por los instrumentos y procesadores ópticos de los telescopios de dónde se obtiene la información, incluyendo lecturas en el espectro óptico así como en el infrarrojo o ultravioleta (identificados con ciertas longitudes de onda particulares), también se incorporarán etiquetas relativas a luminosidad, velocidad de rotación, nomenclatura, etc.

En el modelo de grafos las relaciones entre los nodos son de crucial importancia para los caminos y son fundamentales para la detección de patrones. Pero en el ámbito bajo estudio hay que reconocer que no existe una única característica que relaciones dos estrellas y sus lecturas asociadas.

Existen multitud de parámetros que pueden indicar relaciones, tales como la distancia entre las estrellas, la luminosidad, el tamaño, las diferentes lecturas espectrales, su período de rotación, etc. Como es impracticable establecer relaciones para todas las posibles combinaciones de características, la implementación permitirá generar dichas relaciones a demanda.

Es decir, como paso previo al análisis y detección de patrones, el usuario deberá indicar qué característica o combinación de ellas quiere utilizar para generar las relaciones entre las estrellas. Acto seguido se generarán las relaciones indicadas y el análisis se realizará en base a el grafo resultante con la estructura determinada de acuerdo a las características indicadas por el usuario.

Dicho enfoque se ha utilizado oportunamente en los trabajos publicados por Coutinho et. al 2016[11], un ejemplo de lo cual se presenta en la Figura 2.

Figura 2. Galaxias representadas como grafo etiquetado. Coutinho et. al

3.3. Fuentes y pre-procesamiento de los datos

Los mencionados proyectos de survey astronómicos cuentan con grandes repositorios de información accesible de forma pública, los cuales van a ser utilizados como fuentes de información para cargar la base de datos de análisis.

El estandar de facto para el intercambio de información astronómica se basa en el formato de archivos FITS (Flexible Image Transport System[18]) el cual será utilizado como base para todas las rutinas de importación de datos.

ONgDB cuenta con la posibilidad de importar archivos de datos de forma masiva, lo cual es muy necesario debido a que los archivos fuente generalmente cuentan con volúmenes de aproximadamente medio millón de registros. Para ello utiliza archivos separados por comas con un formato particular

Es necesario realizar un procesamiento previo de la información para que la misma se encuentre en un formato adecuado para su importación masiva en la base de datos. Para ello se propone un desarrollo en lenguaje Python, el cual provee las librerías AstroPy[27] que son adecuadas para la tarea de lectura y parseo de los archivos FITS y permite gestionar con cierta facilidad los archivos CSV de destino. El procedimiento completo está descripto en la Figura 3.

Figura 3. Pre-procesamiento e importación de datos

3.4. El almacenamiento seleccionado

Para la implementación del modelo de datos se optó por la utilización de ONgDB (https://www.graphfoundation.org/projects/ongdb/), una alternativa completamente Open Source a Neo4J (https://neo4j.com/), la que quizá sea la base de datos de grafos comercial más difundida.

Se optó por este producto debido a que reúne algunas características muy valiosas para el proyecto:

- Representa de forma nativa las características del modelo de grafo etiquetado mencionado anteriormente
- Se basa en un producto comercial de probada calidad y tecnología actualizada
- Dispone de una licencia no restrictiva que permite utilizarlo de manera libre y gratuita en instituciones educativas y como parte de proyectos de investigación

- No tiene restricciones en cuanto a características avanzadas tales como
  - Transacciones ACID
  - Replicación
  - Monitoreo

4. Conclusiones y trabajos futuros

En vista a los resultados obtenidos durante la elaboración del presente trabajo se considera que las bases de datos de grafos son un mecanismo válioso para el almacenamiento de información astronómica, brindando un esquema flexible pero que a la vez puede realizar consultas de manera eficiente en grandes volúmenes de datos.

Los grafos etiquetados permiten reflejar con exactitud los datos astronómicos, permitiendo asimismo el crecimiento y adecuación de las estructuras de acuerdo a las necesidades que surjan a partir de los archivos de datos, tarea ésta que es dificultosa en el caso de utilizar bases de datos relacionales.

En posteriores etapas durante la evolución del plan de tesis se prevé optimizar los mecanismos de importación de información astronómica y brindar a los usuarios algunas herramientas intuitivas para realizar tanto las tareas de importación de datos como la definición de los criterios para la generación de las relaciones a demanda.

Se desarrollará un sistema de consulta que permita indicar los parámetros a utilizar en las búsquedas y se implementará un módulo que obtenga algunos indicadores estadísticos relevantes de acuerdo a los datos almacenados.

Posteriormente se analizarán algunos modelos de patrones existentes en otras disciplinas y se implementarán tests para utilizar dichos patrones en un ámbito astronómico. La finalidad de ésta actividad es determinar si existen patrones preexistentes que tengan aplicación en un ámbito para el que no han sido diseñados. Un especial interés tienen los algoritmos de redes de mundo pequeño[22] y los algoritmos de redes sociales[10].

Referencias

1. Alonso Romero, L. y Calonge Cano, T. Redes neuronales y reconocimiento de patrones (2001).
2. Ball, N. M. y Brunner, R. J. Data mining and machine learning in astronomy. *International Journal of Modern Physics D* **19,** 1049-1106 (2010).
3. Baume, G. *y col.* Basic parameters of three star clusters in the Small Magellanic Cloud: Kron 11, Kron 63 and NGC 121. *Monthly Notices of the Royal Astronomical Society* **390,** 1683-1690 (2008).
4. Bondy, J. A. y Murty, U. S. R. *Graph theory with applications* (Citeseer, 1976).
5. Borne, K. D. Astroinformatics: a 21st century approach to astronomy. *arXiv preprint arXiv:0909.3892* (2009).

6. Borne, K. D. en *Next Generation of Data Mining* 114-137 (Chapman y Hall/CRC, 2008).
7. Bundy, K. *y col.* Overview of the SDSS-IV MaNGA survey: mapping nearby galaxies at Apache Point observatory. *The Astrophysical Journal* **798,** 7 (2014).
8. Bunke, H. y Allermann, G. Inexact graph matching for structural pattern recognition. *Pattern Recognition Letters* **1,** 245-253 (1983).
9. Burges, C. J. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery* **2,** 121-167 (1998).
10. Carrington, P. J., Scott, J. y Wasserman, S. *Models and methods in social network analysis* (Cambridge university press, 2005).
11. Coutinho, B. *y col.* The network behind the cosmic web. *arXiv preprint arXiv:1604.03236* (2016).
12. Dalcanton, J. J. *y col.* The ACS nearby galaxy survey treasury. *The Astrophysical Journal Supplement Series* **183,** 67 (2009).
13. Devijver, P. A. y Kittler, J. *Pattern recognition theory and applications* (Springer Science & Business Media, 2012).
14. Fan, W. *Graph pattern matching revised for social network analysis* en *Proceedings of the 15th International Conference on Database Theory* (2012), 8-21.
15. Frinchaboy, P. M. *y col.* en *Star Clusters in the Era of Large Surveys* 31-38 (Springer, 2012).
16. Fukunaga, K. *Introduction to statistical pattern recognition* (Elsevier, 2013).
17. Gilmore, G. *y col.* The Gaia-ESO public spectroscopic survey. *The Messenger* **147,** 25-31 (2012).
18. Hanisch, R. J. *y col.* Definition of the flexible image transport system (FITS). *Astronomy & Astrophysics* **376,** 359-380 (2001).
19. Ivezić, Ž. *y col.* Astrometry with digital sky surveys: from SDSS to LSST. *Proceedings of the International Astronomical Union* **3,** 537-543 (2007).
20. Kim, H. Y. y Giacomantone, J. O. *A new technique to obtain clear statistical parametric map by applying anisotropic diffusion to fMRI* en *Image Processing, 2005. ICIP 2005. IEEE International Conference on* **3** (2005), III-724.
21. Kim, H. Y., Giacomantone, J. y Cho, Z. H. Robust anisotropic diffusion to produce enhanced statistical parametric map from noisy fMRI. *Computer Vision and Image Understanding* **99,** 435-452 (2005).
22. Kleinberg, J. M. Navigation in a small world. *Nature* **406,** 845 (2000).
23. Meyer-Baese, A. y Meyer-Baese, A. *Pattern recognition for medical imaging* (Academic Press, 2004).
24. Minniti, D. *y col.* VISTA Variables in the Via Lactea (VVV): The public ESO near-IR variability survey of the Milky Way. *New Astronomy* **15,** 433-443 (2010).
25. Pavlidis, T. *Structural pattern recognition* (Springer, 2013).
26. Robinson, I., Webber, J. y Eifrem, E. *Graph databases: new opportunities for connected data* (.º'Reilly Media, Inc.", 2015).

27. Robitaille, T. P. *y col.* Astropy: A community Python package for astronomy. *Astronomy & Astrophysics* **558,** A33 (2013).
28. Schmeja, S. Identifying star clusters in a field: A comparison of different algorithms. *Astronomische Nachrichten* **332,** 172-184 (2011).
29. V., B., H, B. y A., F. *Data Science and Classification* (Springer, 2006).
30. Van Harmelen, F., Lifschitz, V. y Porter, B. *Handbook of knowledge representation* (Elsevier, 2008).
31. Van Steen, M. Graph theory and complex networks. *An introduction* **144** (2010).
32. Watanabe, S. *Pattern recognition: human and mechanical* (John Wiley & Sons, Inc., 1985).