UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

REGLAS DE ASOCIACIÓN PARA LÍNEAS MOLECULARES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL EN COMPUTACIÓN

NICOLÁS MARTÍN MIRANDA CASTILLO

PROFESOR GUÍA:
GUILLERMO CABRERA VIVES

MIEMBROS DE LA COMISIÓN:

Este trabajo ha sido parcialmente financiado por .

SANTIAGO DE CHILE
DICIEMBRE 2014

Resumen

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Una dedicatoria corta. Por ejemplo, A los creadores de U-Campus

Agradecimientos

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Índice general

Índice de figuras

Introducción

[...] we may in time ascertain the mean temperature of heavenly bodies, but I regard this order of facts as for ever excluded from our recognition. We can never learn their internal constitution [...]

Auguste Comte, *Astronomy, Ch. I: General View*, 1835

En los últimos tiempos, y en gran parte debido al explosivo desarrollo tecnológico, han surgido numerosos campos en los cuales se ha requerido el uso de procesamiento masivo de datos e inteligencia computacional con el fin de automatizar y auxiliar el proceso de generación de nuevo conocimiento. La astronomía es, sin lugar a dudas, uno de ellos. Esto se debe, en parte, al explosivo desarrollo de nuevas tecnologías que ponen al alcance de la comunidad científica una cantidad nunca antes vista de datos; los cuales tienen el potencial de contener invaluable información sobre el universo, su composición, estructura, origen y destino.

Un claro ejemplo de esto lo constituye el *Atacama Large Millimiter/sub-millimiter Array (ALMA)*, un interferómetro radio-astronómico que consiste en un arreglo de 66 antenas que observan el espacio en las bandas milimétricas y submilimétricas del espectro electromagnético. Ubicado en el desierto de Atacama, en el norte de Chile, es parte de uno de los proyectos científicos más importantes del último tiempo a nivel nacional; en el cual se ha hecho uso de tecnologías de punta por parte de investigadores, ingenieros y técnicos expertos en computación de alto rendimiento, redes de fibra óptica, Machine Learning, minería de datos, entre otros.

La tecnología involucrada en el proyecto *ALMA* ha permitido, entre otras cosas, obtener datos de alta resolución provenientes de distintas fuentes u objetos del espacio observable desde la tierra, cuya posición en el cielo es cuantificada mediante coordenadas celestes. La radiación electromagnética emitida por estos objetos, en bandas de frecuencia de radio, son captadas por el arreglo de antenas y posteriormente procesadas por equipos de alta capacidad con el fin de obtener los espectros electromagnéticos correspondientes. Estos, a su vez pueden ser analizados directamente o utilizarse para generar imágenes de alta calidad de regiones del espacio en diversos rangos de frecuencia.

Si bien el caso de *ALMA* es notable por las características particulares de las observaciones

en frecuencias de radio y por las oportunidades tecnológicas que involucra, el analizar y extraer información a partir de radiación electromagnética, en diversos rangos de su espectro, es parte primordial de la labor astronómica en todos sus campos.

Parte principal de la importancia de estos espectros de radiación electromagnética es que dan información valiosa sobre la composición química de los objetos de los que esta proviene; ya sean estrellas, galaxias u otros muchos tipos de estructuras celestiales. Esto se debe a que los átomos que componen estos objetos emiten o absorben una mayor cantidad de energía en frecuencias muy específicas, tal y como lo predice la teoría cuántica subyacente. Por lo tanto, un espectro en particular tendrá puntos más altos (picos) en ciertas frecuencias dependiendo de los elementos químicos de los que está compuesto el objeto del que proviene.

La detección de lineas espectrales es un problema de interés en sí, y que puede llegar a ser muy complejo dependiendo de en qué bandas de frecuencia se esté trabajando. Sin embargo, se puede seguir obteniendo información valiosa de los objetos observados a partir de estas líneas ya detectadas. Esto incluye potencialmente respuestas a preguntas como: ¿de qué forma se relacionan ciertos tipos de líneas entre sí? ¿Existe una mayor correlación de presencia de líneas de ciertos isótopos o moléculas en particular? ¿Hay una mayor presencia de líneas de ciertas especies en algunos objetos que en otros? ¿Qué nos dice esto de la composición de los objetos y de su química subyacente?

Capítulo 1

Marco Teórico

Gran parte de los datos obtenidos desde ALMA son guardados en estructuras de datos llamadas cubos de datos tipo ALMA (o *ALMA Data Cubes*), que contienen información de distintos puntos de observación del cielo a distintas frecuencias. Los espectros de frecuencia son una forma de representar la intensidad de la radiación electromagnética, recibida desde un punto del espacio, en un cierto rango de frecuencias. Estos contienen puntos altos de intensidad en ciertas frecuencias en las cuales se sabe que una cierta molécula conocida efectúa una transición cuántica. Por lo tanto, mediante reconocer e identificar estos puntos altos, o *peaks*, se puede saber las transiciones moleculares que ocurrieron en el objeto del que proviene la radiación electromagnética. Como, a su vez, se sabe de antemano a qué frecuencia específicamente se realizan las transiciones cuánticas de moléculas conocidas, puede inferirse cuáles son las moléculas presentes en el objeto de origen.

A su vez, los cubos de datos tipo ALMA, como estructura de datos, contienen valores indexados en tres coordenadas. Dos de las coordenadas son espaciales, y corresponden al equivalente a una imagen normal de dos dimensiones, en el sentido que describen puntos del cielo (o del espacio observable desde la tierra). La tercera coordenada corresponde al rango de frecuencias en el que se está detectando radiación electromagnética. Por lo tanto, si se fijan las dos coordenadas espaciales (se fija un punto en el espacio) y se extraen todos los valores en la tercera coordenada de aquel punto, se obtiene el espectro de frecuencias observado en ese punto del espacio.

Los cubos de datos tipo ALMA, por tanto, contienen información de los espectros de frecuencia observados en todos los puntos de un sector dado del espacio. Todos los espectros presentes en un cubo se encuentran en un mismo rango de frecuencia. Sin embargo, distintos cubos de datos pueden tener observaciones hechas en distintos rangos de frecuencia entre sí.

A partir de ALMA se generan enormes cantidades de datos, los cuales, debido a su gran tamaño, necesariamente deben procesarse por parte de sistemas automatizados de extracción y análisis con el fin de facilitar a los investigadores el extraer información útil a partir de estos. La mayoría de estas herramientas se encuentran dentro de las áreas de investigación de la minería de datos y Machine Learning; disciplinas de la computación que han tenido un gran auge en el último tiempo.

1.1. Reglas de asociación

Dentro del área del aprendizaje computacional automatizado, o Data Mining, existe una técnica que ha sido ampliamente utilizada e investigada desde su concepción. Se trata del aprendizaje mediante reglas de asociación, o *Association Rule Learning (ARL)*; la cual se creó con el fin de identificar relaciones entre los productos preferidos por los consumidores de distintos sistemas de punto de venta, como supermercados, tiendas de venta al detalle, etc. La intuición es que, si se posee una base de datos con transacciones, donde cada una de estas posee un cierto conjunto de ítems que un cliente en particular ha comprado, con la ayuda de un algoritmo pueden encontrarse una serie de reglas que indiquen relaciones entre las compras de ciertos ítems en particular. Un ejemplo de regla (bastante intuitiva, por lo demás) sería: "En el 90 % de las transacciones en que se compró pan y mantequilla también se compró leche". Los algoritmos de ARL permiten obtener relaciones simples, como la del ejemplo, y otras mucho más difíciles de deducir por otros medios.

Formalmente, se considera un conjunto de variables $\mathcal{X} = \{X_j\}_{j=1}^{p}$. Usualmente, estas variables se consideran como binarias: $X_j \in \{0, 1\}$. Se define, entonces un conjunto de N *transacciones* $D = \{t_\mathrm{i}\}_{\mathrm{i}=1}^{N}$, donde $t_\mathrm{i} = \{x_{\mathrm{i},j}\}_{j=1}^{p}$, y

$$x_{\mathrm{i},j} = \begin{cases} 0 & \text{si es ítem } j \text{ es parte de la transacción i} \\ 1 & \text{de lo contrario.} \end{cases}$$

El objetivo principal del análisis mediante reglas de asociación es obtener aquellos valores de variables conjuntas $(X_1.X_2, \ldots, X_p)$ que aparezcan de manera más frecuente en el conjunto de datos.

A partir de éstos se generan reglas de asociación, de la forma:

$$I \Rightarrow X_j \Big| c$$

donde $I \subset \mathcal{X}$, $X_j \notin I$ y c es la *confianza de la regla*, o la razón entre el número de transacciones en D que contienen a $I \cup X_j$ y el número de transacciones que contienen a I. A su vez, la razón entre el número de transacciones que contienen a $I \cup X_j$ y el número total de transacciones se conoce como el *soporte* de la regla de asociación.

Capítulo 2

Especificación del Problema

Capítulo 3

Descripción de la Solución

Capítulo 4

Validación de la Solución

Conclusión

Bibliografía

[1] Atacama large millimeter/submillimeter array (alma). `http://www.almaobservatory.org`, 2014. online, accesed July 2014.

[2] matplotlib: python plotting. `http://www.matplotlib.org`, 2014. online, accesed July 2014.

[3] Numpy – numpy. `http://www.sdss.org`, 2014. online, accesed July 2014.

[4] Welcome to python.org. `http://www.python.org`, 2014. online, accesed July 2014.

[5] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM, 1993.

[6] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, pages 3–14. IEEE, 1995.

[7] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[8] Ferenc Bodon. A fast apriori implementation. In *Proceedings of the IEEE ICDM workshop on frequent itemset mining implementations (FIMI'03)*, volume 90, 2010.

[9] Sergey Brin, Rajeev Motwani, and Craig Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *ACM SIGMOD Record*, volume 26, pages 265–276. ACM, 1997.

[10] Sergey Brin, Rajeev Motwani, Jeffrey D Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In *ACM SIGMOD Record*, volume 26, pages 255–264. ACM, 1997.

[11] Chun Hing Cai, Ada Wai-Chee Fu, CH Cheng, and WW Kwong. Mining association rules with weighted items. In *Database Engineering and Applications Symposium, 1998. Proceedings. IDEAS'98. International*, pages 68–77. IEEE, 1998.

[12] Pedro Carmona-Saez, Monica Chagoyen, Francisco Tirado, Jose M Carazo, and Alberto

Pascual-Montano. Genecodis: a web-based tool for finding significant concurrent annotations in gene lists. *Genome biology*, 8(1):R3, 2007.

[13] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15, 2009.

[14] R Chaves, JM Górriz, J Ramírez, IA Illán, D Salas-Gonzalez, and M Gómez-Río. Efficient mining of association rules for the early diagnosis of alzheimer's disease. *Physics in medicine and biology*, 56(18):6047, 2011.

[15] Edith Cohen, Amos Fiat, and Haim Kaplan. Associative search in peer to peer networks: Harnessing latent semantics. *Computer Networks*, 51(8):1861–1881, 2007.

[16] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296. ACM, 2010.

[17] Luc Dehaspe, Hannu Toivonen, and Ross D King. Finding frequent substructures in chemical compounds. In *KDD*, volume 98, page 1998, 1998.

[18] Cristian Estan, Stefan Savage, and George Varghese. Automatically inferring patterns of resource consumption in network traffic. In *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 137–148. ACM, 2003.

[19] Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, and Johannes Gehrke. Privacy preserving mining of association rules. *Information Systems*, 29(4):343–364, 2004.

[20] Paolo Ferragina and Antonio Gulli. A personalized search engine based on web-snippet hierarchical clustering. *Software: Practice and Experience*, 38(2):189–225, 2008.

[21] Gabriel Ghinita, Panos Kalnis, Ali Khoshgozaran, Cyrus Shahabi, and Kian-Lee Tan. Private queries in location based services: anonymizers are not necessary. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 121–132. ACM, 2008.

[22] Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86, 2007.

[23] Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1):53–87, 2004.

[24] Farkhund Iqbal, Hamad Binsalleeh, Benjamin Fung, and Mourad Debbabi. A unified data mining solution for authorship analysis in anonymous textual communications. *Information Sciences*, 231:98–112, 2013.

[25] Murat Karabatak and M Cevdet Ince. An expert system for detection of breast cancer based on association rules and neural network. *Expert Systems with Applications*, 36(2):3465–3469, 2009.

[26] Georgi Kobilarov, Tom Scott, Yves Raimond, Silver Oliver, Chris Sizemore, Michael Smethurst, Christian Bizer, and Robert Lee. Media meets semantic web–how the bbc uses dbpedia and linked data to make connections. In *The semantic web: research and applications*, pages 723–737. Springer, 2009.

[27] Stefan Kramer, Luc De Raedt, and Christoph Helma. Molecular feature mining in hiv data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 136–143. ACM, 2001.

[28] Wenke Lee and Salvatore J Stolfo. A framework for constructing features and models for intrusion detection systems. *ACM transactions on Information and system security (TiSSEC)*, 3(4):227–261, 2000.

[29] Xin Li, Lei Guo, and Yihong Eric Zhao. Tag-based social interest discovery. In *Proceedings of the 17th international conference on World Wide Web*, pages 675–684. ACM, 2008.

[30] Jong Soo Park, Ming-Syan Chen, and Philip S Yu. *An effective hash-based algorithm for mining association rules*, volume 24. ACM, 1995.

[31] Animesh Patcha and Jung-Min Park. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, 51(12):3448–3470, 2007.

[32] Anthony J Remijan and Andrew J Markwick-Kemper. Splatalogue: Database for astronomical spectroscopy. 2008.

[33] Cristóbal Romero and Sebastian Ventura. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1):135–146, 2007.

[34] Cristóbal Romero, Sebastián Ventura, and Enrique García. Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, 51(1):368–384, 2008.

[35] Ashok Savasere, Edward Robert Omiecinski, and Shamkant B Navathe. An efficient algorithm for mining association rules in large databases. 1995.

[36] Petr Škoda and Jaroslav Vážnỳ. Searching of new emission-line stars using the astroinformatics approach. *arXiv preprint arXiv:1112.2775*, 2011.

[37] Ramakrishnan Srikant and Rakesh Agrawal. Mining generalized association rules. In *VLDB*, volume 95, pages 407–419, 1995.

[38] Ramakrishnan Srikant and Rakesh Agrawal. Mining quantitative association rules in large relational tables. In *ACM SIGMOD Record*, volume 25, pages 1–12. ACM, 1996.

[39] Wei Wang, Jiong Yang, and Philip S Yu. Efficient mining of weighted association rules (war). In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 270–274. ACM, 2000.

[40] Donald G York, J Adelman, John E Anderson Jr, Scott F Anderson, James Annis, Neta A Bahcall, JA Bakken, Robert Barkhouser, Steven Bastian, Eileen Berman, et al. The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579, 2000.

[41] Mohammed Javeed Zaki and Mitsunori Ogihara. Theoretical foundations of association rules. In *3rd ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 71–78. Citeseer, 1998.