No leas para contradecir o refutar, ni para creer o dar por bueno, ni para buscar materia de conversación o de discurso, sino para considerar y ponderar lo que lees.

FRANCIS BACON

Los expertos dicen que dos cosas determinan dónde estará usted dentro de cinco años a partir de ahora: los libros que lee y las personas con las que se asocia.

Bibliografía

[1] Aggarwal, C. C., Sun, Z. & Yu, P. S. (1998). *Online algorithms for finding profile association rules*. Proceedings of the 1998 ACM CIKM 7th International Conference on Information and Knowledge Management. Bethesda, Maryland, USA, November 3-7, 1998, pp. 86-95

Artículo dedicado a la extracción de reglas de asociación con atributos numéricos en el que se propone la utilización de un árbol S para mantener un índice multidimensional.

[2] Aggarwal, C. C. & Yu, P. S. (1998a). *A New Framework for Itemset Generation*. Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Seattle, Washington, June 1-3, 1998, pp. 18-24.

En este artículo se realiza una crítica del modelo clásico empleado en la obtención de reglas de asociación y se propone sustituir los itemsets frecuentes por itemsets "fuertemente colectivos" para conseguir un proceso de obtención de reglas más eficiente y productivo. El modelo empleado trata de eliminar la generación de reglas espúreas (aquéllas que no aportan nada nuevo) y evitar la no obtención de reglas potencialmente interesantes debida al valor de *MinSupport*, el cual, en ocasiones, ha de fijarse demasiado alto para evitar una explosión combinatoria en la generación de reglas.

[3] Aggarwal, C. C. & Yu, P. S. (1998b). *Mining Large Itemsets for Association Rules*. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 1998.

Repaso general sobre los métodos de obtención de reglas de asociación (itemsets, para ser precisos) en el que se propone el uso de la "fuerza colectiva" *[collective strength]* para caracterizar los itemsets involucrados en las reglas de asociación.

[4] Agrawal, R., Imielinski, T. & Swami, A. (1993). *Mining association rules between sets of items in large databases*. Proceedings of the 1993 ACM

SIGMOD International Conference on Management of Data, Washington, D.C., May 26-28, 1993, pp. 207-216

Primer artículo donde se presenta el concepto de regla de asociación y se propone el algoritmo AIS.

[5] Agrawal, R. & Shafer, J.C. (1996). *Parallel Mining of Association Rules*. IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, pp. 962-969, December 1996.

Trabajo en el que se expone cómo implementar el algoritmo Apriori en un multiprocesador sin memoria compartida realizando la comunicación mediante paso de mensajes (con MPI), lo que es aplicable también a un cluster de estaciones de trabajo conectadas a través de una red de interconexión de alta capacidad. Se proponen tres alternativas: *Count Distribution*, *Data Distribution* y *Candidate Distribution*.

[6] Agrawal, R. & Shim, K. (1996). *Developing Tightly-Coupled Applications on IBM DB2/CS Relational Database System: Methodology and Experience*, Second International Conference on Knowledge Discovery in Databases and Data Mining, Portland, Oregon, August 1996, pp. 287-290.

En esta ponencia, resumen de un informe técnico de IBM, se presenta la implementación del algoritmo Apriori de extracción de reglas de asociación utilizando capacidades propias del sistema gestor de bases de datos relacionales de IBM.

[7] Agrawal, R. & Srikant, R. (1994). *Fast Algorithms for Mining Association Rules*. VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile, pp. 487-499

Uno de los artículos más importantes sobre reglas de asociación. En él se presentan los algoritmos *Apriori* y *AprioriTID*, así como su combinación *AprioriHybrid*, para la obtención de todas las reglas de asociación existentes en una base de datos de transacciones. En el artículo se muestra cómo esta familia de algoritmos mejora los algoritmos anteriores (AIS y SETM).

[8] Ali, K., Manganaris, S. & Srikant, R. (1997). *Partial Classification using Association Rules*. Proceedings of the 3rd International Conference on Knowledge Discovery in Databases and Data Mining, August 14-17, 1997, Newport Beach, California, USA, pp. 115-118

Se expone la posibilidad de utilizar reglas de asociación para construir modelos de clasificación parcial que, si bien no siempre consiguen una clasificación precisa, pueden ser útiles para describir clases individuales. Como medida de la importancia que tiene cada regla de asociación a la hora de clasificar se utiliza el "riesgo relativo".

[9] Andersen, A., Blair, G., Goebel, V., Karlsen, R., Stabell-Kulo, T. & Yu, W. (2001). *Arctic Beans: Configurable and reconfigurable enterprise component architectures*. IEEE Distributed Systems Online, Vol. 2, No. 7.

En este artículo se presenta el proyecto Arctic Beans, que pretende dotar de mayor flexibilidad a los sistemas basados en componentes utilizados en la actualidad (como COM+, EJB o CORBA). Su idea es incluir mecanismos que le permitan al sistema configurarse, reconfigurarse y evolucionar de forma semi-autónoma.

[10] Anderson, D.P. & Kubiatowicz, J. (2002). *The Worldwide Computer*, Scientific American, March 2002.

Interesante artículo de los promotores del proyecto SETI@Home que discute los servicios que serían necesarios para crear la infraestructura necesaria para la construcción de un supercomputador virtual que incluyese la capacidad de todos los sistemas conectados a Internet: ISOS [*Internet-Scale Operating System*].

[11] Atkinton, C., Bayer, J., Bunse, C., Kamsties, E., Laitenberger, O., Laqau, R, Muthig, D., Paech, B., Wüst, J. & Zettel, J. (2002). *Component-based product line engineering with UML*. Addison-Wesley Component Software Series. ISBN 0-201-73791-4.

Libro en el que se describe KobrA, un método de desarrollo de software basado en componentes, el cual tiene sus raíces en los métodos dirigidos a objetos, Cleanroom, OPEN y otras técnicas basadas en componentes y orientadas a la creación de líneas de productos.

[12] Aumann, Y. & Lindell, Y. (1999). *A statistical theory for quantitative association rules*. Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 15-18, 1999, San Diego, California, USA, pp. 261-270

Trabajo dedicado a la obtención de reglas de asociación con atributos numéricos en el cual se obtiene la distribución de los valores de los atributos numéricos correspondiente a los itemsets formados por atributos categóricos (que han de obtenerse en primer lugar) para encontrar perfiles que indiquen distribuciones representativas.

[13] Bayardo Jr., R. J. (1997). *Brute-Force Mining of High-Confidence Classification Rules*. En KDD-97, Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, August 14-17, 1997, Newport Beach, California, USA, pp. 123-126.

Se presentan una serie de técnicas de poda que permiten optimizar el proceso de extracción de reglas de asociación destinadas a ser utilizadas para construir clasificadores.

[14] Bayardo Jr., R. J. (1998). *Efficiently mining long patterns from databases*. Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA, pp. 85-93.

En este artículo se propone una forma alternativa de enfrentarse al problema de obtener el conjunto de itemsets frecuentes a la que se utiliza en los algoritmos derivados de *Apriori*. Tales algoritmos no son adecuados cuando hay k-itemsets frecuentes con k elevado, ya que todos sus subconjuntos (2^k) son, a su vez, frecuentes y como tales han de ser tratados. El algoritmo propuesto, *Max-Miner*, extrae eficientemente sólo aquéllos itemsets relevantes que no estén incluidos en otros itemsets relevantes.

[15] Bentley, J. (2000). *Programming Pearls*, 2nd edition. ACM Press / Addison-Wesley, ISBN 0-201-65788-0.

Uno de los pocos libros de Informática que puede considerarse una auténtica joya, por el ingenio y la perspicacia con la que Jon Bentley escribe sus ensayos. Originalmente, el contenido recopilado en este libro apareció en la columna homónima de *Communications of the ACM*.

[16] Bergholz, A. (2000). *Extending your markup: An XML tutorial*. IEEE Internet Computing, July / August 2000, pp. 74-79.

Breve y útil guía para todo aquél que desee familiarizarse con la sintaxis del lenguaje XML [*eXtensible Markup Language*] y la multitud de estándares que le rodea: DTD [*Document Type Definition*], XML Schema, XSL [*eXtensible Stylesheet Language*], XSLT [*XSL Transformations*]...

[17] Berry, M.J.A. & Linoff, G. (1997). *Data Mining Techniques: for Marketing, Sales, and Customer Support*. John Wiley and Sons, 1997.

Un libro destinado a ejecutivos y gerentes para que éstos se familiaricen con distintas técnicas de *Data Mining* existentes, su uso y sus limitaciones. El libro abarca desde el análisis de las transacciones comerciales [*basket data analysis*] hasta la utilización de árboles de decisión, métodos de agrupamiento, redes neuronales y algoritmos genéticos.

[18] Berzal, F., Cubero, J. C., Cuenca, F. & Martín-Bautista, M. J. (2001). *On the quest for easy-to-understand splitting rules*, pendiente de publicación en Data & Knowledge Engineering.

Trabajo en el que se proponen dos criterios de división alternativos (MAXDIF y GG) para la construcción de árboles de decisión. Ambos criterios consiguen resultados comparables a los de cualquier otra regla de división, si bien su complejidad es menor, con lo cual se facilita la comprensión del proceso de construcción del árbol de decisión por parte del usuario final de un sistema de extracción de conocimiento.

[19] Berzal, F., Cubero, J. C., Marín, N. & Serrano, J. M. (2001). *TBAR: An efficient method for association rule mining in relational databases*, Data & Knowledge Engineering, 37 (2001), pp. 47-64.

Artículo donde se presenta el algoritmo $\overline{T}$ (*TBAR*), el cual utiliza una estructura de datos basada en un árbol de enumeración de subconjuntos que permite mejorar las prestaciones de *Apriori* en la obtención de itemsets frecuentes.

[20] Bow, S.-T. (1991). *Pattern Recognition and Image Processing*. Marcel Dekker, 1991. ISBN 0-8247-8583-5.

Libro de texto de Reconocimiento de Formas en el que aparece descrito el algoritmo de agrupamiento basado en grafos citado en la sección 5.1.

[21] Bradshaw, J.M., Greaves, M., Holmback, H., Karygiannis, T., Jansen, W., Suri, N. & Wong, A. (1999). *Agents for the masses?*. IEEE Intelligent Systems, March / April 1999, pp. 53-63.

Informe que delinea las distintas líneas de investigación relacionadas con el desarrollo de sistemas distribuidos con agentes inteligentes. Se hace hincapié en los mecanismos de comunicación entre agentes y en la gestión de sistemas multiagente con el objetivo de simplificar y potenciar el desarrollo de tales sistemas.

[22] Brassard, G. & Bratley, P. (1997). *Fundamentos de algoritmia*. Prentice-Hall, ISBN 84-89660-00-X.

Libro de texto de Teoría de Algoritmos que incluye, entre otros, un capítulo dedicado a distintos algoritmos de exploración de grafos, como la "vuelta atrás" utilizada en TBAR para recorrer el árbol de itemsets y generar un conjunto de reglas de asociación.

[23] Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, California, USA, 1984.

El libro de CART, un clásico en la amplia literatura dedicada a la construcción de árboles de decisión.

[24] Brin, S., Motwani, R., Ullman, J. D. & Tsur, S. (1997). *Dynamic Itemset Counting and Implication Rules for Market Basket Data*. Proceedings of the ACM SIGMOD international conference on Management of Data, May 11 - 15, 1997, Tucson, Arizona, USA, pp. 255-264

Artículo en el que se presenta el algoritmo *DIC [Dynamic Itemset Counting]*. Este algoritmo, derivado de *Apriori*, reduce el número de veces que se ha de recorrer la base de datos para obtener sus itemsets frecuentes. Además, sus autores proponen el uso de "reglas de implicación" basadas en una medida de convicción como alternativa al uso de la confianza en las reglas de asociación.

[25] Butte, T. (2002). *Technologies for the development of agent-based distributed applications*. ACM Crossroads, 8:3, Spring 2002, pp. 8-15.

Análisis de los requisitos necesarios para la implementación de sistemas multiagente en entornos distribuidos en el que se incluye una discusión sobre las "facilidades"que ofrece Java para el desarrollo de este tipo de aplicaciones.

[26] Canavan, J.E. (2001). *Fundamentals of Network Security*. Artech House, 2001. ISBN 1-58053-176-8.

En este libro se analizan distintas vulnerabilidades existentes en los sistemas informáticos distribuidos, se describen algunas amenazas que se pueden presentar en forma de ataques y se describen los distintos mecanismos de seguridad que se pueden emplear para prevenirlos y neutralizarlos (p.ej. técnicas criptográficas de protección de datos).

[27] Chan, P.K. (1989). *Inductive learning with BCT*. Proceedings of the 6th International Workshop on Machine Learning, Ithaca, NY, June 130 - July 2.

En este artículo se propone BCT [*Binary Classification Tree*], un híbrido de ID3 [129] y CN2 [35] que construye árboles binarios en los que cada nodo contiene una expresión booleana definida potencialmente sobre varios atributos.

[28] Chandrasekaran, B., Josephson, J.R. & Benjamins, V.R. (1999). *What are Ontologies, and why do we need them?*. IEEE Intelligent Systems, January/February 1999, pp. 20-26.

Buena introducción al tema de las ontologías, una línea de investigación de moda en Inteligencia Artificial: la creación de teorías acerca de los tipos de objetos existentes, sus propiedades y las relaciones existentes entre ellos, así como su evolución temporal. El proyecto CYC de Lenat y la base de datos Wordnet son ejemplos destacables en este ámbito.

[29] Chaudhuri, S. & Dayal, U. (1997). *An Overview of Data Warehousing and OLAP Technology*. ACM SIGMOD Record, March 1997.

Excelente artículo que describe la arquitectura de un sistema OLAP y analiza algunas cuestiones relativas a su diseño.

[30] Chaudhuri, S., Fayyad, U. & Bernhardt, J. (1999). *Scalable Classification over SQL Databases*. IEEE: Proceedings of the 15th International Conference on Data Engineering, Sydney, Australia, March 23-26, 1999.

Ponencia en la que se propone la utilización de una arquitectura multicapa para construir clasificadores a partir de datos almacenados en bases de datos relacionales. En la arquitectura propuesta por los investigadores de Microsoft, existiría una capa intermedia de software entre la base de datos y el algoritmo de construcción de árboles de decisión que se encargaría de enviar a este último información resumida sobre los datos almacenados en la base de datos.

[31] Cheung, D.W., Han, J., Ng, V.T. & Wong, C.Y. (1996). *Maintenance of Discovered Association Rules in Large Databases: An Incremental Updating Technique*, Proceedings of the Twelfth International Conference on Data Engineering, New Orleans, Louisiana, February 1996, pp. 106-114.

Ponencia que trata del mantenimiento de un conjunto de reglas de asociación dada una serie de modificaciones sobre la base de datos de la que se obtuvieron las reglas, lo que se conoce como extracción incremental de reglas de asociación.

[32] Cheung, D.W., Ng, V.T. & Tam, B.W. (1996). *Maintenance of Discovered Knowledge: A Case in Multi-level Association Rules*, KDD'96, Second International Conference on Knowledge Discovery and Data Mining, Oregon, August 1996, pp. 307-310.

Continuación del artículo anterior [31] en el cual se extienden los algoritmos incrementales de extracción de reglas de asociación para que puedan trabajar con reglas de asociación generalizadas, que son reglas de asociación construidas utilizando jerarquías de conceptos.

[33] Clark, D. (1999). *Service with a (smart) smile: networks Jini-style*. IEEE Intelligent Systems, May / June 1999, pp.81-83.

Es este artículo se describen las características básicas de Jini, una tecnología que facilita el desarrollo de sistemas distribuidos dinámicos. También se comentan el origen de las ideas en que se basa Jini (el proyecto Linda de la Universidad de Yale) y se mencionan sus principales competidores (productos de funcionalidad similar ofrecidos por otras empresas rivales de Sun Microsystems).

[34] Clark, P. & Boswell, R. (1991). *Rule induction with CN2: Some Recent Improvements*. In Y. Kodratoff, editor, Machine Learning - EWSL-91, Berlin, 1991, Springer-Verlag, pp. 151-163

Se proponen dos mejoras sobre el algoritmo CN2 básico: el uso de una estimación laplaciana del error como función de evaluación alternativa y la obtención de un conjunto de reglas no ordenado. Los resultados obtenidos por este CN2 mejorado se comparan con C4.5.

[35] Clark, P. & Nibblett, T. (1989). *The CN2 Induction Algorithm*. Machine Learning Journal, Kluwer Academic Publishers, 3(4) pp. 261-183.

Se expone el algoritmo CN2, que fue desarrollado con el objetivo de conseguir un método eficiente de aprendizaje inductivo para la obtención de reglas de producción en presencia de ruido o de información incompleta. En este trabajo, CN2 se compara con ID3 y AQ.

[36] Clarke, I. (1999). *A distributed decentralises information storage and retrieval system*. University of Edinburg, Division of Informatics, 1999.

Informe en el que se describe la posible implementación de un sistema distribuido de almacenamiento y recuperación de información que no necesita un elemento central de control o administración. El sistema propuesto, implementado en FreeNet (http://freenetproject.org/), permite publicar información de forma anónima y evitar cualquier tipo de censura (se mantienen copias redundantes de los documentos para evitar su posible desaparición).

[37] Codd, E.F., Codd, S.B. & Salley, C.T. (1998). *Providing OLAP to User-Analysts: An IT Mandate*, Hyperion Solutions Corporation, 1998.

Informe de Codd y sus asociados en los que se analiza la evolución de los sistemas OLAP como complemento de los sistemas OLTP tradicionales. Aunque resulta interesante su lectura, hay que mantener una perspectiva crítica respecto a lo que en él afirma (no en vano, es un informe pagado por una casa comercial).

[38] Cohen, W. (1995). *Fast effective rule induction*. Proc. 12th International Conference on Machine Learning, Morgan Kaufmann, 1995, pp. 115-123.

En esta ponencia se presenta el algoritmo RIPPERk, *Repeated Incremental Pruning to Produce Error Reduction*. Este algoritmo constituye una mejora sobre el algoritmo IREP y permite obtener listas de decisión más compactas mejorando la precisión de los clasificadores IREP.

[39] Cortijo Bon, F. (1999). Apuntes de *Reconocimiento de Formas*, E.T.S. Ingeniería Informática, Universidad de Granada.

Esta asignatura, opcional en los estudios de Ingeniería Informática, incluye varios temas relacionados con el contenido de esta memoria. Entre otros temas, en ella se estudian distintos tipos de clasificadores y se analizan varios métodos de agrupamiento.

[40] Coyle, F.P. (2002). *XML, Web Services and the changing face of distributed computing*, Ubiquity, ACM IT Magazine and Forum, Volume 3, Issue 10, April 23-29, 2002.

Artículo en el que se analizan las tendencias actuales relativas a la implementación de sistemas distribuidos, en los cuales se tiende a utilizar protocolos independientes del lenguaje de programación, del sistema operativo y del protocolo de transporte.

[41] Curbera, F., Duftlet, M., Khalaf, R., Nagy, W., Mukhi, N. & Weerawarana, S. (2002). *Unraveling the Web Services Web: An introduction to SOAP, WSDL, and UDDI*. IEEE Internet Computing, March / April 2002, pp. 86-93.

Como su título indica, este artículo repasa las tecnologías en que se basan los servicios web: SOAP como protocolo de comunicación basado en XML, WSDL como lenguaje universal de descripción de servicios y UDDI como especificación de un registro centralizado de servicios a modo de directorio.

[42] Czajkowski, G. & von Eicken, T. (1998). *JRes: a resource accounting interface for Java*. Proceedings of the conference on Object-oriented programming, systems, languages, and applications (OOPSLA'98), Vancouver, British Columbia, Canada, 1998, pp. 21-35.

En este artículo se propone una extensión de la máquina virtual Java estándar, denominada JRes, que permita monitorizar y controlar el uso de memoria, el tiempo de CPU y los recursos de entrada/salida (conexiones de red, ancho de banda...). Para ello se establecen límites sobre el uso de recursos que pueden hacer las distintas hebras de una aplicación multihebra.

[43] Czajkowski, G., Mayr, T., Seshadri, P. & von Eicken, T. (1999). *Resource Control for Java Database Extensions*. 5th USENIX Conference on Object-Oriented Technologies and Systems (COOTS '99), San Diego, California, USA, May 3–7, 1999.

En esta ponencia se explora el uso de JRes para monitorizar el uso de un sistema de bases de datos, detectando posibles ataques distribuidos, monitorizando el uso de recursos de cada usuario y tomando medidas que puedan aprovecharse en la optimización de consultas.

[44] Davis, M. (2001). *Struts, an open-source MVC implementation*, IBM developerWorks, February 2001.

Breve artículo en el que se describe la arquitectura de Struts, que forma parte del proyecto Jakarta, un proyecto amparado por la fundación Apache. Struts implementa el modelo MVC en Java con servlets y JSP.

[45] Díaz, A. Glover, F., Ghaziri, H.M., González, J.L., Laguna, M., Moscato, P. & Tseng, F.T. (1996). *Optimización heurística y redes neuronales*. Editorial Paraninfo, ISBN 84-283-2264-4.

En este libro se analizan distintas técnicas de búsqueda heurística, entre las que se encuentran el enfriamiento simulado (al que se denomina "recocido simulado" en este libro), los algoritmos genéticos, la búsqueda tabú y GRASP [*Greedy Randomized Adaptive Search Procedure*].

[46] Domingos, P. (1996). *Linear-time rule induction*. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), 1996, pp. 96-101

Artículo en el que se describe el algoritmo CWS, que mejora propuestas anteriores como CN2.

[47] Domingos, P. (1998). *Occam's Two Razors: The Sharp and the Blunt*. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), August 27-31, 1998, New York City, USA, pp. 37-43

Crítica al uso del Principio de Economía de Occam como garantía de que la simplicidad de un modelo implica su mayor precisión (si bien la simplicidad como objetivo sí es apropiada en muchas ocasiones, para facilitar la comprensibilidad de los modelos construidos).

[48] Domingos, P. (1999). *The Role of Occam's Razor in Knowledge Discovery*. Data Mining and Knowledge Discovery Volume 3, Number 4, December 1999, pp. 409-425

Muchos sistemas de aprendizaje inductivo incorporan una preferencia explícita por modelos simples (la Navaja de Occam), si bien este criterio no siempre se utiliza correctamente. Este artículo expone que su uso continuado puede llegar a impedir la consecución de resultados interesantes en KDD.

[49] Dong, G. & Li, J. (1999). *Efficient mining of emerging patterns: discovering trends and differences*. Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 15-18, 1999, San Diego, CA USA, pp. 43-52.

En este artículo se propone el uso de EPs *[Emerging Patterns]*, itemsets cuyo soporte se ve incrementado significativamente de un conjunto de datos a otro.

[50] Dong, G., Zhang, X., Wong, L. & Li, J. (1999). *CAEP: Classification by Aggregating Emerging Patterns*. Proceedings of the Second International Conference on Discovery Science, Tokyo, Japan, 1999, pp. 30-42.

En esta ponencia se presenta CAEP, un clasificador que se construye a partir de EPs *[Emerging Patterns]* y consigue excelentes resultados cuando se compara con C4.5 o CBA.

[51] Dougherty, J., Kohavi, R., and Sahami, M. (1995). *Supervised and unsupervised discretization of continuous features*. Proceedings of the 12th International Conference on Machine Learning, Los Altos, CA, Morgan Kaufmann, 1995, pp. 194–202.

Este artículo repasa distintos métodos de discretización utilizados en Inteligencia Artificial. Los autores categorizan distintas propuestas en función de si son métodos supervisados o no supervisados y del uso global o local que se les suele dar.

[52] Dubois, D. and Prade, H. (1993). *Fuzzy sets and probability: misunderstandings, bridges and gaps*. Proceedings of the Second IEEE Conference on Fuzzy Systems, 1993, pp. 1059–1068.

Breve artículo en el que se intenta establecer un enlace entre la Teoría de la Probabilidad y la Teoría de los Conjuntos Difusos.

[53] Dreyfus, H. L. (1994). *What Computers Still Can't Do. A Critique of Artificial Reason*. The MIT Press, fourth printing, 1994.

Libro muy recomendable repleto de comentarios y críticas acerca de los supuestos logros de la Inteligencia Artificial (y de los que investigan en el tema).

[54] Duda, R.O. & Hart, P.E. (1973). *Pattern Classification and Scene Analysis*, John Wiley & Sons, ISBN 0471223611.

El libro de texto clásico de Reconocimiento de Formas. Wiley ha editado recientemente una versión revisada y extendida del libro (Richard O. Duda, Peter E. Hart, David G. Stork: *Pattern Classification (2nd Edition)*, 2000, ISBN 0471056693) en la que se mantiene un capítulo dedicado al aprendizaje no supervisado.

[55] Dykstra, J. (2002). *Software verification and validation with Destiny: A parallel approach to automated theorem proving*. ACM Crossroads, issue 8.3, Spring 2002, pp. 23-27.

En este artículo se describe Destiny, un sistema paralelo de demostración de teoremas con una arquitectura bastante peculiar: centralizada desde el punto de vista del control de la carga del sistema, en anillo desde el punto de vista de los nodos encargados de procesar datos.

[56] Elder IV, J.F. (1995). *Heuristic search for model structure: the benefits of restraining greed*. AI & Statistics - 95, 5th International Workshop on Artificial Intelligence and Statistics, Ft. Lauderdale, Florida, 3-6 January, pp. 199-210.

Ponencia en la que se propone un algoritmo de construcción de árboles de decisión en el que la bondad de una partición no se mide por la pureza de los hijos resultantes, sino por la de los 'nietos'.

[57] Fayad, M. & Schmidt, D.C., eds. (1997). *Object-oriented application frameworks*, Communications of the ACM, October 1997, pp. 32ss.

Sección especial de la revista mensual de la ACM dedicada al desarrollo de sistemas basados en componentes utilizando técnicas de orientación a objetos.

[58] Fayyad, U.M. & Irani, K.B. (1993). *Multi-interval discretization of continuous-valued attributes for classification learning*. Proceedings of the 13th International Joint Conference on Artificial Intelligence, pp. 1022-1027.

Trabajo pionero en la construcción de árboles de decisión n-arios con atributos continuos. Tras demostrar que para obtener el umbral óptimo basta con evaluar los puntos de corte entre casos de distintas clases, propone un método supervisado de discretización basado en el Principio MDL de Rissanen.

[59] Fayyad, U.M., Piatetsky-Shapiro, G. & Smyth, P. (1996). *The KDD process for extracting useful knowledge from volumes of data*, Communications of the ACM, November 1996, pp. 27-34.

Este artículo ofrece una visión general de aquello a lo que nos referimos al hablar de KDD, revisa algunos temas relacionados y concluye con una enumeración de los desafíos a los que hemos de enfrentarnos: gran volumen de datos, información incompleta o redundante, diseño de técnicas de interacción con el usuario, desarrollo de algoritmos incrementales...

[60] Feldman, R., Amir, A., Auman, Y., Zilberstien A. & Hirsh, H. (1997).*Incremental Algorithms for Association Generation*, Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining. En "KDD: Techniques and Applications", H. Lu et al. eds., World Scientific, Singapur, pp. 227-240.

Artículo en el que se decriben distintas técnicas incrementales de extracción de reglas de asociación.

[61] Flammia, G. (2001). *Peer-to-peer is not for everyone*. IEEE Intelligent Systems, May / June 2001, pp. 78-79.

Breve artículo en el que se acotan las posibles aplicaciones que el modelo P2P pueda llegar a tener en el desarrollo futuro, defendiendo la vigencia del modelo cliente/servidor. A diferencia de [10], el autor no espera que los sistemas P2P provoquen un cambio radical en el diseño de sistemas distribuidos.

[62] Fowler, M. (1997). *Analysis patterns: Reusable object models*. Addison-Wesley, ISBN 0-201-89542-0.

Un libro sobre patrones de diseño de utilidad en el análisis orientado a objetos de distintas aplicaciones comunes. En cierto modo, el contenido de este libro se complementa con las monografías de Hay [77] y Gamma et al. [65].

[63] Freitas, A. A. (2000). *Understanding the Crucial Differences between Classification and Discovery of Association Rules - A Position Paper*. SIGKDD Explorations, 2:1, July 2000, pp. 65-69.

Trabajo en el que se marcan claramente las diferencias conceptuales existentes entre clasificación y obtención de reglas de asociación.

[64] Fürnkranz, J., and Widmer, F. (1994). *Incremental reduced error pruning*. In Machine Learning: Proceedings of the 11th Annual Conference, New Brunswick, New Jersey, Morgan Kaufmann, 1994.

Ponencia que describe el algoritmo de inducción de listas de decisión IREP, *Incremental Reduced Error Pruning*.

[65] Gamma, E., Helm, R., Johnson, R. & Vlissides, J. (1995). *Design Patterns*, Addison-Wesley, ISBN: 0-201-633612.

El libro de "la Banda de los Cuatro", basado en la tesis doctoral de Erich Gamma, que marcó el comienzo del interés que hoy suscitan los patrones de diseño (en general, el estudio de buenas soluciones a problemas de desarrollo de software como complemento del tradicional estudio de las técnicas que permiten obtenerlas).

[66] Gause, D.C. & Wweinberg, G.M. (1989). *Exploring Requirements: Quality Before Design*, Dorset House, September 1989, ISBN: 0932633137.

Interesante libro de Don Gause y Jerry Weinberg en el que se tratan distintas técnicas útiles a la hora de realizar el análisis de requisitos de un sistema; esto es, cómo descubrir qué es lo que hay que hacer para avanzar en el árbol de decisión que conduce a la resolución de un problema.

[67] Gehrke, J., Ganti, V., Ramakrishnan, R. & Loh, W.-Y. (1999a). *BOAT - Optimistic Decision Tree Construction*. Proceedings of the 1999 ACM SIGMOD international conference on Management of Data, May 31 - June 3, 1999, Philadelphia, PA USA, pp. 169-180

Artículo en el que se presenta BOAT, un algoritmo de la familia TDIDT que permite la construcción eficiente, escalable, e incluso incremental de árboles de decisión al seleccionar un subconjunto de datos que se utiliza para generar un árbol que posteriormente se refina.

[68] Gehrke, J., Loh, W.-Y. & Ramakrishnan, R. (1999b). *Classification and regression: money can grow on trees*. Tutorial notes for ACM SIGKDD 1999 international conference on Knowledge Discovery and Data Mining, August 15-18, 1999, San Diego, California, USA, pp. 1-73

Tutorial de KDD'99 en el que se ofrece un repaso de distintos algoritmos de construcción de árboles de decisión.

[69] Gehrke, J, Ramakrishnan, R. & Ganti, V. (2000). *RainForest - A Framework for Fast Decision Tree Construction of Large Datasets*. Data Mining and Knowledge Discovery, Vol. 4, No. 2/3, pp. 127-162.

En este artículo se presenta una familia de algoritmos TDIDT denominada *RainForest*. Se propone un método general de construcción eficiente de árboles de decisión separando sus propiedades de escalabilidad de los demás factores que influyen en la calidad de los árboles construidos.

[70] Gilb, T. (1988). *Principles of Software Engineering Management*. Addison-Wesley, ISBN 0-201-19246-2.

En esta obra se defiende un enfoque incremental en el desarrollo de software, haciendo especial hincapié en el establecimiento de objetivos concretos cuantificables. Aunque su lectura es algo tediosa, este libro está repleto de buenos consejos e ideas prácticas.

[71] Giordana, A. & Neri, F. (1996). *Search-intensive concept induction*. Evolutionary Computation 3(4):375-416, Massachusetts Institute of Technology.

En este artículo se describe REGAL, que emplea un algoritmo genético distribuido para inducir descripciones de conceptos (esto es, reglas de clasificación) utilizando Lógica de Primer Orden.

[72] Gong, L., ed., (2002). *Peer-to-peer networks in action*. IEEE Internet Computing, January / February 2002, pp. 37ss.

Sección especial dedicada a una de las tecnologías más prometedoras de las que han aparecido en los últimos años, que ha dado lugar a multitud de proyectos interesantes, tanto desde el punto de vista teórico como desde el punto de vista práctico. FreeNet, Gnutella y JXTA Search se encuentran entre los sistemas analizados en los artículos que componen esta sección de IEEE Internet Computing.

[73] Glymour, C., Madigan, D., Pregibon, D. & Smyth, P. (1996). *Statistical Inference and Data Mining*. Communications of the ACM, November 1996, pp. 35-41.

Este artículo trata de lo que la Estadística puede aportar a las técnicas de Data Mining: básicamente, la evaluación de las hipótesis generadas y de los resultados obtenidos.

[74] Han, J., Pei, J. & Yin, Y. (2000). *Mining Frequent Patterns without Candidate Generation*. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 15 - 18, 2000, Dallas, TX USA, pp. 1-12

Se presenta el *FP-Tree [Frequent-Pattern Tree]*, que permite representar una base de datos transaccional de una forma compacta. En una primera pasada por el conjunto de datos, se obtienen los items frecuentes y se establece un orden entre ellos (de más a menos frecuente). En un segundo recorrido secuencial de los datos se construye el árbol. A partír de él se pueden obtener todos los itemsets frecuentes de forma recursiva.

[75] Han, J. & Kamber, M. (2001). *Data Mining: Concepts and Techniques*. Academic Press, Morgan Kauffman Publishers, 2001.

Libro de texto muy recomendable que recoge en sus páginas muchas de las técnicas utilizadas para resolver problemas de *Data Mining*, desde el uso de *data warehouses* en aplicaciones OLAP hasta la extracción de reglas de asociación o la construcción de distintos tipos de clasificadores.

[76] Han, J.L. & Plank, A.W. (1996). *Background for Association Rules and Cost Estimate of Selected Mining Algorithms*. CIKM '96, Proceedings of the Fifth International Conference on Information and Knowledge Management, November 12 - 16, 1996, Rockville, Maryland, USA, pp. 73-80

En este artículo se intenta comparar, desde un punto de vista estadístico, el coste computacional de distintos algoritmos para la obtención de reglas de asociación (Apriori, AprioriTid, AprioriHybrid, OCD, SETM y DHP) así como estudiar sus propiedades de escalabilidad.

[77] Hay, D.C. (1995). *Data Model Patterns*. Dorset House Publishing, ISBN 0-932633-29-3.

Libro muy interesante dedicado al modelado de datos en el que se describen distintos patrones que pueden ser de utilidad como punto de partida en el diseño de la base de datos de un sistema de información. Más allá de su interés académico, resulta útil para familiarizarse con muchos conceptos y procesos que aparecen durante el desarrollo de cualquier aplicación de gestión empresarial.

[78] Hedberg, S.R. (1999). *Data Mining takes off at the speed of the Web*. IEEE Intelligent Systems, November / December 1999, pp. 35-37.

En este informe se comenta la importancia económica que han adquirido las técnicas de Data Mining en el competitivo mundo empresarial: un mercado de 800 millones de dólares en 2000, con más de 200 empresas ofreciendo soluciones que incorporan técnicas de Data Mining, según META Group.

[79] Herrera, F. (1999). Apuntes de *Bioinformática*, E.T.S. Ingeniería Informática, Universidad de Granada.

Esta asignatura, opcional en los estudios de Ingeniería Informática, incluye un tema dedicado a la utilización de algoritmos genéticos en la construcción de sistemas clasificadores.

[80] Hidber, C. (1999). *Online Association Rule Mining*. Proceedings of the 1999 ACM SIGMOD international conference on Management of Data, May 31 - June 3, 1999, Philadelphia, PA, USA, pp. 145-156

Ponencia en la que se presenta *CARMA [Continuous Association Rule Mining Algorithm]*, un derivado de *Apriori* y *DIC* que sólo necesita recorrer dos veces la base de datos para obtener todos los itemsets frecuentes.

[81] Hipp, J., Güntzer, U. & Nakhaeizadeh, G. (2000). *Algorithms for Association Rule Mining - A General Survey and Comparison*. SIGKDD Explorations, Volume 2, Issue 1, June 2000, pp. 58-64

Estudio comparativo y análisis de las similitudes existentes entre los distintos algoritmos que se han propuesto para la extracción de reglas de asociación en el que se resaltan los aspectos comunes a todos ellos.

[82] Ho, K.M., and Scott, P.D. (1997). *Zeta: A global method for discretization of continuous variables*. 3rd International Conferemce on Knowledge Discovery and Data Mining (KDD'97), Newport Beach, CA, AAAI Press, pp. 191-194.

Aquí se presenta un método supervisado de discretización basado en Zeta, una medida del grado de asociación existente entre los valores de dos atributos categóricos.

[83] Holte, R.C. (1993). *Very simple classification rules perform well on most commonly used datasets*. Machine Learning, 11:63-90.

Artículo en el que se presentó el discretizador 1R (*One Rule*), un sencillo método supervisado de discretización.

[84] Houtsma, M. & Swami, A. (1993). *Set-oriented mining for association rules*. IBM Research Report RJ9567, IBM Almaden Research Center, San Jose, California, October 1993.

Informe en el que se presentó el algoritmo SETM, un algoritmo equivalente a AIS [4] orientado hacia su implementación en SQL sobre bases de datos relacionales.

[85] Hussain, F., Liu, H., Tan, C.L., and Dash, M. (1999). *Discretization: An enabling technique*. The National University of Singapore, School of Computing, TRC6/99, June 1999.

Este informe repasa distintos métodos de discretización (todos ellos de tipo jerárquico) y propone una taxonomía que permite clasificarlos.

[86] Imielinski, T. & Mannila, H. (1996). *A Database Perspective on Knowledge Discovery*. Communications of the ACM, November 1996, pp. 58-64.

En este artículo se analizan los métodos empleados en Data Mining desde la perspectiva de las bases de datos. Se ponen de manifiesto las limitaciones de SQL a la hora de construir aplicaciones de Data Mining y se expone la necesidad de idear lenguajes de consulta más potentes. Tal como dijo C.J. Date, "El modelo relacional representa el lenguaje ensamblador de los sistemas modernos (y futuros) de bases de datos".

[87] Inmon, W.H. (1996). *The Data Warehouse and Data Mining*. Communications of the ACM, November 1996, pp. 49-50.

En este breve artículo se hace hincapié en que la calidad de los datos recopilados y de la forma en que se almacenan en un *data warehouse* es esencial para obtener buenos resultados al aplicar técnicas de *Data Mining*.

[88] Joshi, M.V., Agarwal, R.C., and Kumar, V. (2001). *Mining needles in a haystack: Classifying rare classes via two-phase rule induction*. Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data, Santa Barbara, California, 2001, pp. 91-101.

Los autores proponen un algoritmo de inducción de listas de decisión adecuado para problemas de decisión binarios en los cuales una de las clases es mucho menos frecuente que la otra, lo cual puede provocar que algoritmos como RIPPER o C4.5 no resulten adecuados.

[89] Juristo, N., Windl, H. & Constantine, L., eds. (2001). *Introducing Usability*, IEEE Software Issue on Usability Engineering, Vol. 18, No. 1, January/February 2001, pp. 20ss.

Sección dedicada a la difusión de técnicas que, utilizadas durante el proceso de desarrollo de software, permitan mejorar la usabilidad de los sistemas de información: facilidad de aprendizaje, eficiencia, prevención de errores y satisfacción del usuario.

[90] Kiczales, G., Hilsdale, E., Hugunin, J., Kersten, M., Palm, J. & Griswold, W.G. (2001). *Getting started with AspectJ*. Communications of the ACM, Vol. 44, No. 10, October 2001, pp. 59-65.

En este artículo se presenta AspectJ, una extensión "orientada a aspectos" del popular lenguaje de programación Java. Básicamente, la "orientación a aspectos" pretende mejorar la cohesión de las distintas facetas de un producto software, facilitando su implementación y posterior mantenimiento. Para lograrlo, se aplican a los lenguajes de programación estándar técnicas similares a las utilizadas por los disparadores en las bases de datos relacionales.

[91] Kobryn, C. (2000). *Modeling components and frameworks with UML*, Communications of the ACM, Volume 43, Issue 10, October 2000, pp. 31-38.

Artículo donde se describe el patrón de diseño utilizado por sistemas basados en componentes como J2EE (Sun Microsystems) o .NET (Microsoft Corporation).

[92] Kodratoff, Y. (1988). *Introduction to Machine Learning*. Pitman Publishing, 1988.

Libro sobre *Machine Learning* en el que destaca su tercer apéndice, *"ML in Context"*, en el que se realiza una curiosa analogía entre la educación y el aprendizaje automático.

[93] Kodratoff, Y. (2001). *Comparing Machine Learning and Knowledge Discovery in Databases*. Lecture Notes in Artificial Intelligence, Springer, LNAI 2049, pp. 1-21.

Interesante capítulo de libro que revisa distintas medidas que se pueden utilizar para caracterizar una regla $A \Rightarrow B$ y estimar su validez.

[94] Langley, P. (1996). *Elements of Machine Learning*. Morgan Kaufmann Publishers, 1996.

Libro de texto que pretende ser algo así como una ”tabla periódica”de métodos de aprendizaje e intenta eliminar las fronteras, artificiales en muchas ocasiones, que delimitan los distintos paradigmas existentes: inducción de reglas, aprendizaje analítico, aprendizaje basado en casos, redes neuronales, algoritmos genéticos...

[95] Larsen, G., ed. (2000). *Component-based enterprise frameworks*, Communications of the ACM, October 2000, pp. 24ss.

Sección especial de la revista mensual de la ACM dedicada a la construcción de sistemas basados en componentes y sus aplicaciones comerciales.

[96] Leavitt, N. (2002). *Industry Trends: Data Mining for the corporate masses?*. Computer, IEEE Computer Society, May 2002, pp. 22-24.

En este artículo se expone la situación actual del mercado mundial del *Data Mining*, en el que se destaca la aparición de nuevos estándares y se citan los principales desafíos a los que han de enfrentarse las empresas del sector, para lo cual se hace necesaria la invención de técnicas más eficientes y escalables, la integración con los sistemas de bases de datos existentes y el diseño de aplicaciones más fáciles de usar.

[97] Lee, J.H., Kim, W.Y., Kim, M.H., and Lee, J.L. (1993). *On the evaluation of boolean operators in the extended boolean retrieval framework*. ACM SIGIR’93, Pittsburgh, pp. 291-297

Interesante ponencia en la que se comentan las propiedades de los operadores booleanos que se emplean en distintos modelos (booleanos, obviamente) de Recuperación de Información.

[98] Lee, C.-H. & Shim, D.-G. (1999). *A multistrategy approach to classification learning in databases*. Data & Knowledge Engineering 31, 1999, pp. 67-93

Trabajo en el que se propone un algoritmo híbrido de aprendizaje que combina inducción de reglas para construir un modelo de clasificación parcial con el algoritmo k-NN para clasificar aquellos casos no cubiertos por las reglas obtenidas. También se propone en este artículo la utilización de la divergencia Hellinger para caracterizar las reglas obtenidas como medida de disimilitud que expresa el impacto del antecedente en la distribución del consecuente.

[99] Liu, A. (2001). *J2EE in 2001*, Component Development Strategies, Cutter Information Corp., September 2001.

En este informe se describe la plataforma Java 2 Enterprise Edition de Sun Microsystems y se analizan algunas de sus características más destacadas, incluyendo discusiones acerca de cómo afectan al desarrollo de aplicaciones de gestión en el ámbito empresarial.

[100] Liu, B., Hsu, W. & Ma, Y. (1998). *Integrating Classification and Association Rule Mining*. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), August 27-31, 1998, New York City, USA, pp. 80-86.

Se expone cómo emplear reglas de asociación para construir un clasificador, *CBA [Classification Based on Associations]*, el cual, utilizando una clase por defecto, puede utilizarse como modelo de clasificación completo.

[101] Liu, B., Hu, M. & Hsu, W. (2000a) *Intuitive Representation of Decision Trees Using General Rules and Exceptions*. Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000), July 30 - August 3, 2000, Austin, Texas.

Se construye un clasificador a partir de reglas de asociación utilizando una representación jerárquica que consiste en reglas generales y excepciones a esas reglas (en vez del modelo tradicional en el cual la existencia de demasiadas reglas dificulta la comprensibilidad del modelo construido).

[102] Liu, B., Hu, M. & Hsu, W. (2000b) *Multi-Level Organization and Summarization of the Discovered Rule*. Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 20 - 23, 2000, Boston, MA USA, pp. 208-217

Usando la misma aproximación que en [101], se puede obtener un resumen de la información contenida en un conjunto arbitrario de reglas.

[103] Liu, B., Ma, Y. & Wong, C.K. (2000c). *Improving an Association Rule Based Classifier*. Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2000), September 13-16, 2000, Lyon, France.

Propone mejorar CBA [100] permitiendo la existencia de múltiples umbrales de soporte mínimo para las diferentes clases del problema y recurriendo a algoritmos TDIDT tradicionales cuando no se encuentran reglas lo suficientemente precisas.

[104] Loh, W.-Y. & Shih, Y.-S. (1997). *Split Selection Methods for Classification Trees*. Statistica Sinica, Vol.7, 1997, pp. 815-840

Trabajo en el que se presenta el algoritmo QUEST de construcción de árboles de decisión, que utiliza un criterio de selección no basado en medidas de impureza y no tiende a seleccionar atributos con un grado de ramificación elevado (como sucede en ID3, por ejemplo).

[105] Lopez de Mantaras, R. (1991). *A Distance-Based Attribute Selection Measure for Decision Tree Induction*. Machine Learning, 6, pp. 81-92.

Artículo en el que se propone una normalización de la ganancia de información alternativa al criterio de proporción de ganancia de C4.5. En este caso, la normalización se basa en una métrica de distancia.

[106] Maniatty, W.A. & Zaki, M.J. (2000). *Systems support for scalable Data Mining*. SIGKDD Explorations, December 2000, Volume 2, Number 2, pp. 56-65.

Artículo publicado en un boletín de uno de los grupos de interés especial de la ACM en el cual se comentan los requerimientos de un sistema paralelo de KDD, centrándose especialmente en su subsistema de entrada/salida.

[107] Mannila, H., Toivonen, H. & Verkamo, A.I. (1993): *Improved methods for finding association rules*. University of Helsinky, Department of Computer Science, C-1993-65, December 1993.

Partiendo del algoritmo AIS de Agrawal, Imielinski y Swami, se proponen algunas mejoras. Por ejemplo, calcular $C[k+1]$ a partir de $C[k] \times C[k]$ en vez de usar $L[k] \times L[k]$ para reducir el número de veces que se ha de recorrer la base de datos. Además, se propone podar a priori el conjunto de candidatos $C[k+1]$ generado a partir de $L[k]$, la base del algoritmo *Apriori* desarrollado independientemente en IBM.

[108] Martin, J. K. (1997). *An Exact Probability Metric for Decision Tree Splitting and Stopping*. Machine Learning, 28, pp. 257-291.

Artículo en el que se puede encontrar un estudio exhaustivo sobre distintas reglas de división propuestas para la construcción de árboles de decisión, la correlación entre ellas y resultados experimentales con árboles de decisión binarios.

[109] McConnell, S. (1996). *Rapid Development: Taming wild software schedules*. Microsoft Press, ISBN 1-55615-900-5.

Este libro, que en su día recibió un *Jolt award* (los oscars en el mundo del software), recopila un extenso conjunto de técnicas que pueden ser útiles en el desarrollo de software. Se discuten las virtudes y defectos de cada una de las técnicas y se incluye gran cantidad de datos obtenidos empíricamente relativos al proceso de desarrollo de software. Es memorable el capítulo dedicado a los errores clásicos que se cometen una y otra vez en proyectos de desarrollo de software.

[110] McConnell, S. (1998). *Software Project Survival Guide*. Microsoft Press, ISBN 1-57231-621-7.

Otra obra del editor jefe de IEEE Software en la que se describe un conjunto de técnicas útiles para garantizar, en la medida en que esto es posible, la finalización con éxito de un proyecto de desarrollo de software. Son dignas de mención las listas de comprobación que el autor incluye en cada capítulo para poder evaluar la progresión adecuada de un proyecto.

[111] Mehta, M., Agrawal, R. & Rissanen, J. (1996). *SLIQ: A Fast Scalable Classifier for Data Mining*. Advances in Database Technology - Proceedings of the Fifth International Conference on Extending Database Technology (EDBT'96), Avignon, France, March 25-29, 1996, pp. 18-32

En este trabajo se presenta el algoritmo SLIQ, un algoritmo de construcción de árboles de decisión diseñado específicamente para aplicaciones de Data Mining dentro del proyecto Quest de IBM en el Almaden Research Center de San Jose, California.

[112] Meretakis, D. & Wüthrich, B. (1999). *Extending naïve Bayes classifiers using long itemsets*. Proceedings of the fifth ACM SIGKDD international conference on Knowledge Discovery and Data Mining, August 15-18, 1999, San Diego, CA USA, pp. 165-174

Se propone el algoritmo *LB [Large Bayes]*, que utiliza itemsets frecuentes para mejorar los resultados que se pueden conseguir con el algoritmo Naïve Bayes.

[113] Meyer, B. (2001). *.NET is coming*. IEEE Computer, August 2001, pp. 92-97.

En este artículo, escrito por el creador del lenguaje de programación Eiffel, se analiza la plataforma .NET de Microsoft. A diferencia de la plataforma Java, que es independiente del sistema operativo pero está ligada al lenguaje Java, la plataforma .NET es independiente del lenguaje de programación pero requiere utilizar sistemas operativos de Microsoft (al menos por ahora).

[114] Miller, R. J. & Yang, Y. (1997). *Association rules over interval data*. Proceedings of the ACM SIGMOD Conference on Management of Data, May 11-15, 1997, Tucson, AZ USA, pp. 452-461

Trabajo sobre reglas de asociación con atributos numéricos en el que se encuentra poco adecuada la utilización del soporte y la confianza para caracterizar las reglas obtenidas y se propone el uso de un algoritmo de clustering (BIRCH) para obtener intervalos potencialmente interesantes y una medida del grado de asociación entre los items intervalares obtenidos.

[115] Mitchell, T.M. (1997). *Machine Learning*. McGraw-Hill College Div, ISBN: 0070428077.

Este libro de texto, dedicado enteramente al estudio de distintas técnicas de aprendizaje automático, incluye un capítulo destinado al análisis de distintos algoritmos de aprendizaje de conjuntos de reglas, como, por ejemplo, FOIL.

[116] OMG - The Object Management Group (2000). *OMG Unified Modeling Language Specification*, Version 1.3, First Edition, March 2000, http://www.omg.org/technology/uml/

La especificación completa del Lenguaje Unificado de Modelado, un estándar de facto utilizado en el diseño y análisis de sistemas orientados a objetos.

[117] Othman, O., O'Ryan C. & Schmidt, D. C. (2001). *Strategies for CORBA Middleware-Based Load Balancing*. IEEE Distributed Systems Online, Vol. 2, No. 3, March 2001.

En este artículo se presentan las deficiencias de algunas técnicas comunes de balanceado de carga y se propone una nueva, que los autores han diseñado utilizando características estándar de CORBA y TAO, un ORB concreto.

[118] Özsu, M.T. & Valduriez, P. (1991). *Principles of distributed database systems*. Prentice-Hall International, ISBN 0-13-715681-2.

Posiblemente uno de los mejores libros de texto de bases de datos, al menos a juicio del autor de esta memoria. Incluye un excelente resumen de los conceptos en que se fundamentan las bases de datos relacionales y una gran cantidad de información referente a la implementación de sistemas gestores de bases de datos.

[119] Park, J.S., Chen, M.S. & Yu, P.S. (1995). *An effective hash-based algorithm for mining association rules*. Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, San Jose, California, May 22-25, 1995, pp. 175-186

En esta ponencia se propone el algoritmo para disminuir el número de itemsets candidatos generados en las primeras iteraciones por algoritmos derivados de Apriori. Además, el tamaño de la base de datos que se debe recorrer se va reduciendo paulatinamente en cada iteración del algoritmo (las transacciones en las que ya sabemos que no se encuentra ningún itemset frecuente no hay por qué comprobarlas en las siguientes iteraciones).

[120] Park, J.S., Chen, M.S. & Yu, P.S. (1997). *Using a Hash-Based Method with Transaction Trimmming for Mining Association Rules*. IEEE Transactions on Knowledge and Data Engineering, Vol. 9, No. 5, 1997, pp. 813-825.

Artículo básicamente idéntico a la ponencia de los autores en ACM SIGMOD'95. Como 'novedad', se propone reducir el número de pasadas realizadas por la base de datos generando los candidatos $C[k+1]$ a partir de $C[k] \times C[k]$ en vez de utilizar $L[k] \times L[k]$. Esto es válido si el conjunto de k-itemsets candidatos es lo suficientemente pequeño como para caber en memoria principal.

[121] Patil, A.N. (2001). *Build your own Java-based supercomputer*. IBM developerWorks, April 2001.

Propone la utilización de hebras pseudo-remotas para simplificar el desarrollo de programas paralelos. La planificación de tareas se realiza en una única máquina local, aunque el código correspondiente a las tareas se puede ejecutar en máquinas remotas.

[122] Pazzani, M.J. (2000). *Knowledge discovery from data?*, IEEE Intelligent Systems, March / April 2000, pp. 10-13.

Michael Pazzani realiza una dura crítica de la escasa consideración que solemos tener los que diseñamos sistemas de Data Mining con respecto al usuario final de estos sistemas. Pazzani propone añadir la Psicología Cognitiva a las tres áreas de conocimiento que usualmente se asocian con Data Mining (esto es, Estadística, Bases de Datos e Inteligencia Artificial) con el fin de mejorar la utilidad real de los sistemas de KDD y facilitar la obtención de información útil a partir de grandes conjuntos de datos.

[123] Perry, D.E. & Kaiser, G.E. (1991). *Models of software development environments*. IEEE Transactions on Software Engineering, March 1991, Vol. 17, No. 3, pp. 283-295.

Artículo en el que se propone la caracterización de un sistema de desarrollo de software mediante un modelo general con tres componentes (modelo SMP: estructuras, mecanismos y políticas). A partir de ese modelo básico, se delinean cuatro clases de sistemas en función de su escala utilizando una metáfora sociológica (taxonomía IFCS: individuo, familia, ciudad y estado).

[124] Piatetsky-Shapiro, G. (1999): *The Data-Mining Industry coming of age*, IEEE Intelligent Systems, November / December 1999, pp. 32-34.

Artículo de opinión en el que analizan las aplicaciones y tendencias que pueden marcar la evolución de las técnicas de Data Mining en productos comerciales: creación de estándares, paquetes software como Clementine (SPSS) o Intelligent Miner (IBM) y soluciones verticales para aplicaciones específicas (banca, telecomunicaciones, biotecnología, CRM, comercio electrónico, etc.).

[125] Piatetsky-Shapiro, G., editor (2001): *KDnuggets News*, Vol. 1, No. 11, Item 2, May 29, 2001

Resultados de una encuesta efectuada por *KDnuggets* acerca de las aplicaciones más populares de las técnicas de extracción de conocimiento en bases de datos y *Data Mining*: la banca (17 %) y comercio electrónico (15 %); las telecomunicaciones (11 %); la biología y, en particular, la genética (8 %); la detección de fraude (8 %) y la investigación científica en general (8 %).

[126] Plauger, P.J. (1993). *Programming on Purpose: Essays on Software Design*. PTR Prentice Hall, ISBN 0-13-721374-3.

Excelente libro que recopila algunas de las mejores columnas publicadas por Plauger en la revista *Computer Language*, que hoy sigue publicándose mensualmente bajo el título *Software Development* y mantiene a algunas de las mejores firmas del momento entre sus colaboradores.

[127] Pressman, R.S. (1993). *Ingeniería del Software: Un enfoque práctico*. McGraw-Hill, 3ª edición.

La biblia de la Ingeniería del Software, donde se puede encontrar información sobre casi todo lo relacionado con el proceso de desarrollo de software y la cita con la que se abre esta memoria.

[128] Provost, F. & Kolluri, V. (1999). *A survey of methods for scaling up inductive algorithms*, Data Mining and Knowledge Discovery, volume 3, number 2, pp. 131-169.

Estudio en profundidad de tres estrategias que se pueden seguir para poder aplicar algoritmos de aprendizaje a grandes conjuntos de datos: diseñar un algoritmo rápido (restringiendo el espacio de búsqueda, afinando heurísticas de búsqueda de soluciones, optimizando algoritmos existentes o utilizando paralelismo), dividir el conjunto de datos (uso de muestreo, técnicas incrementales o aprendizaje cooperativo) y utilizar una representación relacional de los datos (p.ej. ontologías: lógica proposicional y jerarquías de conceptos).

[129] Quinlan, J.R. (1986a). *Induction on Decision Trees*. Machine Learning, 1, 1986, pp. 81-106

Uno de los trabajos clásicos en *Machine Learning*. En él se describe el algoritmo ID3, que se enmarca dentro de la familia TDIDT de sistemas de aprendizaje inductivo.

[130] Quinlan, J.R. (1986b). *Learning Decision Tree Classifiers*. ACM Computing Surveys, 28:1, March 1996, pp. 71-72

Breve artículo que ofrece una visión general de la construcción de árboles de decisión para resolver problemas de clasificación.

[131] Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993. ISBN 1-55860-238-0.

El libro del C4.5, uno de los muchos derivados de ID3. No obstante, el algoritmo C4.5 puede considerarse un híbrido entre CART y C4.

[132] Quinlan, J.R. (1996). *Improved use of continuous attributes in C4.5*. Journal of Artificial Intelligence Research, 4:77-90.

Artículo en el que Quinlan propone algunas mejoras sobre el algoritmo C4.5 cuando intervienen atributos de tipo numérico. Se propone un ajuste de la ganancia obtenida en función del número de valores distintos de un atributo (C4.5 Rel.8) y se comenta la posibilidad de utilizar métodos de discretización local al construir el árbol de decisión.

[133] Rasmussen, E. (1992). *Clustering algorithms*. En W.B. Frakes y R. Baeza-Yates (eds.): *Information Retrieval: Data Structures and Algorithms*, Chapter 16.

Capítulo dedicado a los métodos de agrupamiento en uno de los monográficos clásicos dedicados a problemas de recuperación de información.

[134] Rastogi, R. & Shim, K. (2000). *PUBLIC: A Decision Tree Classifier that integrates building and pruning*. Data Mining and Knowledge Discovery, Vol. 4, No. 4, pp. 315-344.

Artículo en el que se propone PUBLIC, un algoritmo de construcción de árboles de decisión que integra la poda del árbol en su construcción: un nodo no se llega a expandir cuando se está construyendo el árbol cuando se determina que se podaría posteriormente durante la post-poda (con la consiguiente mejora computacional que supone ahorrarse la construcción de nodos que después se eliminarían).

[135] Rivest, R.L. (1987). *Learning decision lists*. Machine Learning Journal, vol. 2, no. 3, pp. 229-246, 1987.

El trabajo en el que presenta formalmente el concepto de lista de decisión como método de representación de funciones booleanas.

[136] Sánchez, D. (1999). *Adquisición de relaciones entre atributos en bases de datos relacionales*. Tesis doctoral, Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada, diciembre de 1999.

Tesis doctoral cuya principal aportación consiste en proponer el uso de factores de certeza en vez de utilizar el omnipresente valor de confianza para caracterizar las reglas de asociación extraídas de una base de datos relacional.

[137] Sarawagi, S., Thomas, S. & Agrawal, R. (1998). *Integrating Association Rule Mining with Relational Database Systems: Alternatives and Implications*, SIGMOD 1998, Proceedings of the ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA, pp. 343-354.

En esta ponencia, resumen de un informe realizado en el seno del proyecto Quest de IBM, se discuten las distintas alternativas existentes a la hora de implementar algoritmos de extracción de reglas de asociación en bases de datos relacionales. El abanico de posibilidades abarca desde una implementación débilmente acoplada al DBMS en la cual se utilizan lenguajes de propósito general e interfaces estándar de acceso a bases de datos hasta una implementación no portable fuertemente acoplada al sistema gestor de bases de datos particular que haga uso de interfaces no estándar con el objetivo de obtener un rendimiento máximo.

[138] Segal, R. & Etzioni, O. (1994). *Learning decision lists using homogeneous rules*. AAAI-94, American Association for Artificial Intelligence.

Segal y Etzioni proponen el algoritmo BruteDL, que emplea la fuerza bruta para construir un clasificador basado en reglas. Dado que la precisión de una lista de decisión no es una función directa de la precisión de las reglas que la componen, BruteDL realiza una búsqueda en profundidad para obtener un conjunto de reglas en el cual la interpretación de cada regla no depende de su posición.

[139] Sen, A. (1998). *From DSS to DSP: A taxonomic retrospection*. Communications of the ACM Virtual Extension, Mayo 1998, volumen 41, número 5.

Artículo retrospectivo en el que se analiza la evolución de los sistemas de ayuda a la decisión desde sus orígenes hasta la actualidad, evolución que está muy ligada al progreso de las técnicas de Inteligencia Artificial y de los sistemas de bases de datos.

[140] Sestito, S. & Dillon, T.S. (1994). *Automated Knowledge acquisition*. Sydney, Australia: Prentice Hall, 1994.

Libro sobre técnicas de aprendizaje automático que incluye, entre otros, capítulos dedicados por completo a la construcción de árboles de decisión y a la inducción de reglas utilizando la metodología STAR de Michalski.

[141] Shafer, J.C., Agrawal, R. & Mehta, M. (1996). *SPRINT: A Scalable Parallel Classifier for Data Mining*. VLDB'96, Proceedings of 22th International Conference on Very Large Data Bases, September 3-6, 1996, Mumbai (Bombay), India, pp. 544-555

Artículo sobre SPRINT, un algoritmo eficiente y escalable de construcción de árboles de decisión diseñado para ser paralelizado, lo cual lo hace idóneo para trabajar con grandes bases de datos. Derivado de SLIQ [111], también fue desarrolado dentro del proyecto Quest de IBM.

[142] Shah, M.A., Madden, S., Franklin, M.J. & Hellerstein, J.M. (2001). *Java support for data-intensive systems: Experiences building the Telegraph Dataflow System*. ACM SIGMOD Record, Volume 30, Number 4, December 2001, pp. 103-114.

En este artículo se destacan algunos de los placeres de la programación en Java de aplicaciones de cálculo intensivo, así como algunas de sus limitaciones.

[143] Shih, Y.-S. (1999). *Families of splitting criteria for classification trees*. Statistics and Computing, vol.9, no.4; Oct. 1999; pp. 309-315.

Estudio acerca de las reglas de división utilizadas para construir árboles de decisión binarios.

[144] Shohan, Y. (1999) *What we talk about when we talk about software agents*. IEEE Intelligent Systems, March / April 1999, pp. 28-31.

En esta columna de opinión se critica la propaganda exagerada que existe en torno a la idea de los agentes software, si bien también se resalta la existencia de trabajos de interés que utilizan el término agente en tres ámbitos diferentes: nuevos sistemas expertos (con aplicaciones, por ejemplo, en recuperación de información), sistemas distribuidos (como el descrito en el capítulo 6) y diseño antropomórfico (con el objetivo de obtener máquinas con un comportamiento 'casi humano').

[145] Silverstein, C., Brin, S. & Motwani, R. (1998). *Beyond market baskets: Generalizing association rules to dependence rules*, Data Mining and Knowledge Discovery, 2:39-68.

Artículo en el que se propone la utilización de una medida de interés para caracterizar las reglas de asociación, en vez de utilizar la confianza de la regla.

[146] Srikant, R. & Agrawal, R. (1996). *Mining Quantitative Association Rules in Large Relational Tables*. Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada, June 4-6, 1996, pp. 1-12

En este artículo se plantea el problema de la extracción de reglas de asociación en bases de datos que contengan atributos numéricos y se propone la utilización del método de partición equitativa *[equi-depth partitioning]*.

Para obtener las reglas de asociación se expone una variante del algoritmo Apriori que hace uso de una medida de interés para podar los conjuntos de candidatos.

[147] Skirant, R., Vu, Q. & Agrawal, R. (1997). *Mining Association Rules with Item Constraints*. KDD'97 Proceedings of the 3rd International Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California, August 1997, pp. 67-73.

Este trabajo se centra en la imposición de restricciones al formato de las reglas de asociación y cómo esas restricciones permiten guiar el proceso de extracción de reglas de asociación.

[148] Spell, B. (2000). *Professional Java Programming*, Wrox Press Ltd., December 2000.

Este manual de programación avanzada en Java incluye un capítulo dedicado a la implementación de sistemas distribuidos, comparando el uso de TCP (sockets), CORBA y RMI. Obviamente, RMI es la mejor opción cuando todo el sistema está escrito en Java.

[149] Stang, M. & Whinston, S. (2001). *Enterprise computing with Jini technology*. IT Pro, IEEE Computer Society, January / February 2001, pp. 33-38.

En este artículo se describen las ventajas que ofrece Jini a la hora de construir sistemas distribuidos autoconfigurables. Entre ellas, por ejemplo, se encuentra la posibilidad de distribuir código ejecutable (esto es, se puede dotar de movilidad a los agentes de un sistema multiagente con facilidad).

[150] Taylor, P. C. & Silverman, B. W. (1993). *Block diagrams and splitting criteria for classification trees*. Statistics and Computing, vol. 3, no. 4, pp. 147-161.

Artículo en el que se propone la regla MPI [Mean Posterior Improvement] como alternativa al índice de Gini para construir árboles de decisión binarios.

[151] Tou, J. T., and Gonzalez, R. C. (1974). *Pattern Recognition Principles*. Addison-Wesley, 1974. ISBN 0-201-07587-3.

Libro de texto de Reconocimiento de Formas en el que aparece descrito, con un enrevesado diagrama de flujo de varias páginas, el algoritmo ISODATA citado en la sección 5.1.

[152] Touzet, D., Menaud, J.M., Weis, F., Couderc, P. & Banâtre, M. (2001). *SIDE Surfer: Enriching casual meetings with spontaneous information gathering*. ACM SIGARCH Computer Architecture News, Vol. 29, No.5,

pp. 76-83, December 2001. Ubiquitous Computing and Communication Workshop, PACT (Parallel Architecture and Compilation Techniques) Conference, Barcelona, Septiembre 2001.

En esta ponencia se utiliza una estructura de datos como la de TBAR (capítulo 4) para construir perfiles de usuario que permitan un intercambio de información espontáneo entre dispositivos inalámbricos próximos (tipo PDA).

[153] Ullman, J. (2000). *Data Mining Lecture Notes*. Stanford University CS345 Course on Data Mining, Spring 2000. http://www-db.stanford.edu/∼ullman/mining/mining.html

Apuntes de un curso de Data Mining impartido en la Universidad de Stanford por Jeffrey Ullman, una de las máximas autoridades en el campo de las bases de datos.

[154] Van de Merckt, T. (1993). *Decision trees in numerical attribute spaces*. Proceedings of the 13th International Joint Conference on Artificial Intelligence, pp. 1016-1021.

Artículo en el que se propone la utilización de una medida de contraste para discretizar atributos continuos al construir árboles de decisión. Se basa en la idea de buscar agrupamientos tan distantes como sea posible en el espacio, independientemente de si contienen o no elementos de distintas clases. También se propone una variante supervisada que tiene en cuenta la entropía de la partición resultante.

[155] Vanhelsuwé, L. (1997). *Create your own supercomputer with Java*. JavaWorld, January 1997

Artículo en el que se propone la utilización de Java para realizar tareas de cómputo distribuido utilizando UDP para transferir información de una máquina a otra.

[156] Waltz, D. & Hong S.J., eds. (1999). *Data Mining: A long-term dream*. IEEE Intelligent Systems, November / December 1999, pp. 30ss.

En este número de la revista de IEEE apareció una sección dedicada por completo al creciente interés que despiertan las técnicas de Data Mining, tanto en los centros de investigación como en las empresas, por sus múltiples aplicaciones en biotecnología, compañías aseguradoras, detección de fraude, mantenimiento de aeronaves...

[157] Wang, X., Chen, B., Qian, G. & Ye, F. (2000). *On the optimization of fuzzy decision trees*. Fuzzy Sets and Systems, 112, Elsevier Science B.V., pp. 117-125.

En este artículo se describe la utilización de conjuntos difusos en la construcción de árboles de decisión, dando lugar a árboles en que los caminos de la raíz a las hojas no son mutuamente excluyentes y un caso particular puede corresponder a varias hojas simultáneamente con distintos grados de pertenencia.

[158] Wang, K., Zhou, S. & He, Y. (2000). *Growing decision trees on support-less association rules*. Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 20-23, 2000, Boston, MA USA, pp. 265-269

Trabajo en el que se construye un árbol a partir de reglas de asociación considerando únicamente la confianza de las reglas. El umbral de soporte mínimo se elimina del proceso de extracción de reglas de asociación porque el soporte de una regla no resulta indicativo de su capacidad predictiva.

[159] Wang, K., Zhou, S. & Liew, S.C. (1999). *Building Hierarchical Classifiers Using Class Proximity*. VLDB'99, Proceedings of 25th International Conference on Very Large Data Bases, September 7-10, 1999, Edinburgh, Scotland, UK, pp. 363-374

Se emplea una estrategia similar a la de CBA [100] para construir un clasificador contextual que permita clasificar documentos por temas en jerarquías de conceptos.

[160] Wayne, R. (2002). *Peer (to peer) pressure: it's a good thing*. Software Development, 10:4, April 2002, pp. 38-43.

Artículo en el que se le da un repaso a los sistemas P2P existentes en la actualidad y se analiza su posible evolución en el fututo: "the next big thing?".

[161] Widom, J. (1995). *Research Problems in Data Warehousing*. Proceedings of the 1995 International Conference on Information and Knowledge Management, CIKM'95, November 29 - December 2, 1995, Baltimore, MD, USA, pp. 25ss.

Ponencia en la que se exponen algunos de los problemas que hay que resolver para construir aplicaciones OLAP.

[162] Yeager, W. & Williams, J. (2002). *Secure peer-to-peer networking: The JXTA example*. IT Pro, IEEE Computer Society, March / April 2002, pp. 53-57.

En este artículo se describe el proyecto JXTA haciendo especial énfasis en aspectos relativos a temas de seguridad. La plataforma JXTA para el desarrollo de aplicaciones P2P está basada en XML y es independiente tanto del

lenguaje de programación como del sistema operativo y de los protocolos de red subyacentes.

[163] Zheng, Z. (2000). *Constructing X-of-N Attributes for Decision Tree Learning*. Machine Learning 40(1), July 2000, pp. 35-75

Trabajo en el que se muestra la utilidad del uso simultáneo de varios atributos para ramificar un árbol de decisión.

[164] Zwick, R., Carlstein, E., Budescu, D.V. (1987). *Measures of similarity among fuzzy concepts: A comparative analysis*. International Journal of Approximate Reasoning, 1987, 1:221-242.

Interesante estudio que analiza distintas formas de medir la similitud entre dos entidades desde el punto de vista de la Psicología.

Que otros se jacten de las páginas que han escrito; a mí me enorgullecen las que he leído.

JORGE LUIS BORGES