Índice general

Capítulo 8

Pruebas y resultados

8.1. Rendimiento algoritmos de asociación

El conjunto de datos utilizados en las pruebas pertenecen a las transacciones de uno de los supermercados más importantes del departamento de Nariño (Colombia) durante un periodo determinado. El conjunto de datos contiene 10.757 diferentes productos. Los conjuntos de datos minados con la herramienta TariyKDD se muestran en el cuadro 8.1.

Para cada conjunto de datos se realizó preprocesamiento y transformación de datos con el fin de eliminar los productos repetidos en cada transacción y posteriormente se cargaron las tablas objeto de un modelo simple (i.e. una tabla con esquema Tid, Item) a la estructura de datos DataSet descrito en el capítulo 7 en la sección de implementación.

Cuadro 8.1: Conjuntos de Datos

Nomenclatura	**Numero de Registros**	**Numero de Transaccions**	**Promedio items por transaccion**
BD85KT7	555.123	85.692	7
BD40KT5	194.337	40.256	5
BD10KT10	97.824	10.731	10

Se evaluó el rendimiento de los algoritmos Apriori, FP-Growth y Equipasso, comparando los tiempos de respuesta para diferentes soportes mínimos. Los resultados de la evaluación del tiempo de ejecución de estos algoritmos, aplicados a los

conjuntos de datos BD85KT7, BD40KT5 y BD10KT10, se pueden observar en las figuras 8.1, 8.2 y 8.3 respectivamente.

En general, observando el comportamiento de los algoritmos FP-Growth y Equipasso con los diferentes conjuntos de datos, se puede decir que su rendimiento es similar, contrario al tiempo de ejecución de Apriori, que se ve afectado significativamente a medida que se disminuye el soporte.

Cuadro 8.2: Tiempos de ejecución tabla BD85KT7

BD85KT7			
Soporte (%)	Tiempo (ms)		
	Apriori	FPGrowth	EquipAsso
4.15	750	166	85
4.75	362	162	82
5.35	365	164	83
5.95	365	162	83
6.55	120	159	80

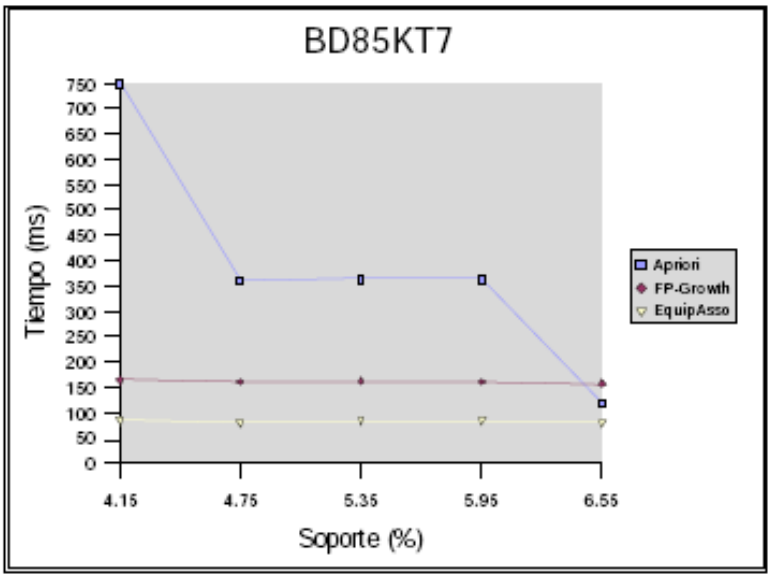

Figura 8.1: Rendimiento BD85KT7

Cuadro 8.3: Tiempos de ejecución tabla BD40KT5

BD40KT5			
Soporte (%)	Tiempo (ms)		
	Apriori	FPGrowth	EquipAsso
1.90	268	66	29
2.00	265	64	29
2.10	132	63	27
2.20	45	61	27
2.30	44	61	27

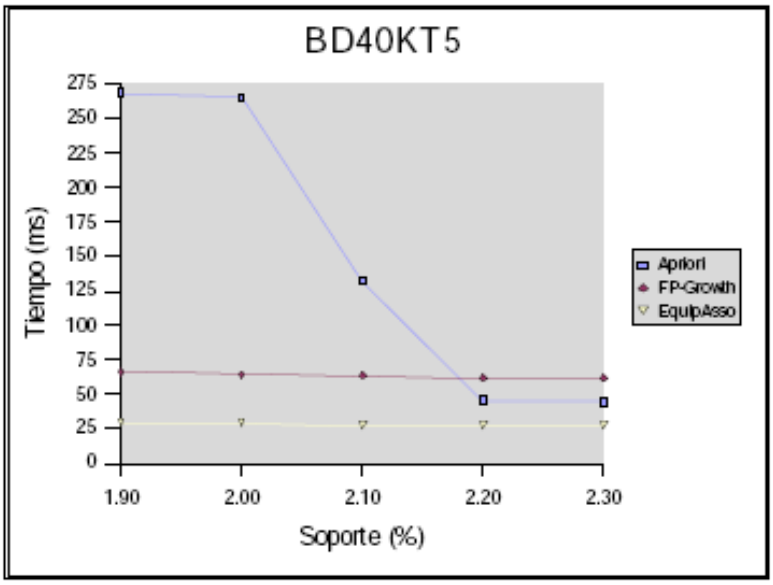

Figura 8.2: Rendimiento BD40KT5

Cuadro 8.4: Tiempos de ejecución tabla BD10KT10

BD10KT10			
Soporte (%)	Tiempo (ms)		
	Apriori	FPGrowth	EquipAsso
3.00	525	28	15
4.00	182	26	14
5.00	105	25	13
6.00	53	24	13
7.00	52	24	13

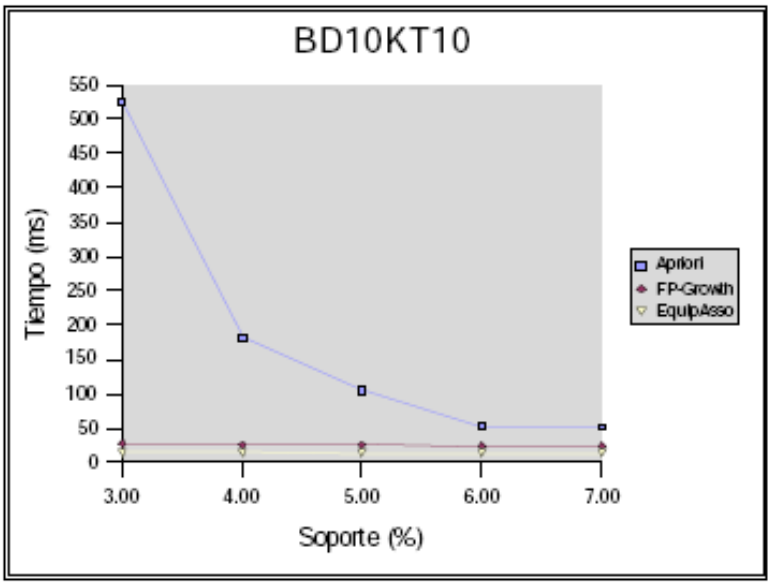

Figura 8.3: Rendimiento BD10KT10

Analizando el tiempo de ejecución de únicamente los algoritmos FP-Growth y EquipAsso (figuras 8.4, 8.5 y 8.6) para los conjuntos de datos BD85KT7, BD40KT5 y BD10KT10. Con soportes más bajos, el comportamiento de estos algoritmos sigue siendo similar.

Cuadro 8.5: Tiempos de ejecución tabla BD85KT7

BD85KT7			
Soporte (%)	Tiempo (ms)		
	Apriori	FPGrowth	EquipAsso
1.00	-	5130	1225
1.50	-	730	270
2.00	-	212	205
2.50	-	202	202
3.00	-	185	187

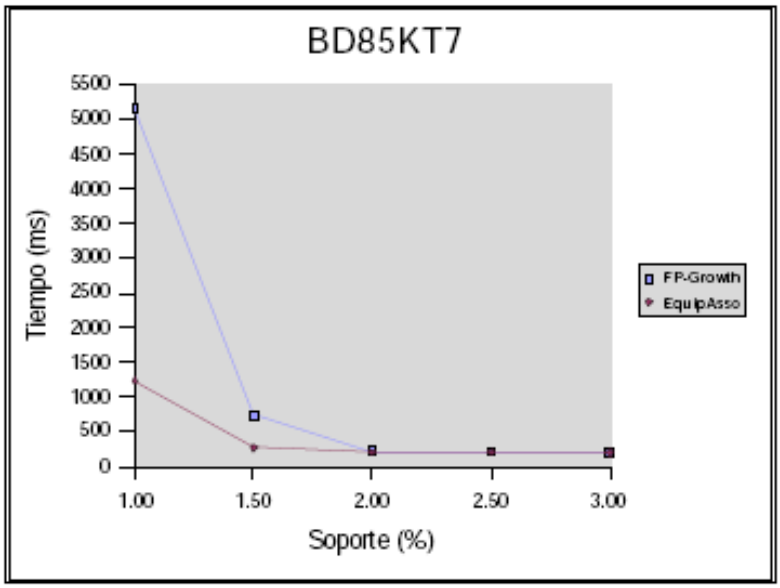

Figura 8.4: Rendimiento BD85KT7

Cuadro 8.6: Tiempos de ejecución tabla BD40KT5

BD40KT5			
Soporte (%)	Tiempo (ms)		
	Apriori	FPGrowth	EquipAsso
0.10	-	965	741
0.20	-	425	290
0.30	-	240	168
0.40	-	156	121
0.50	-	124	105

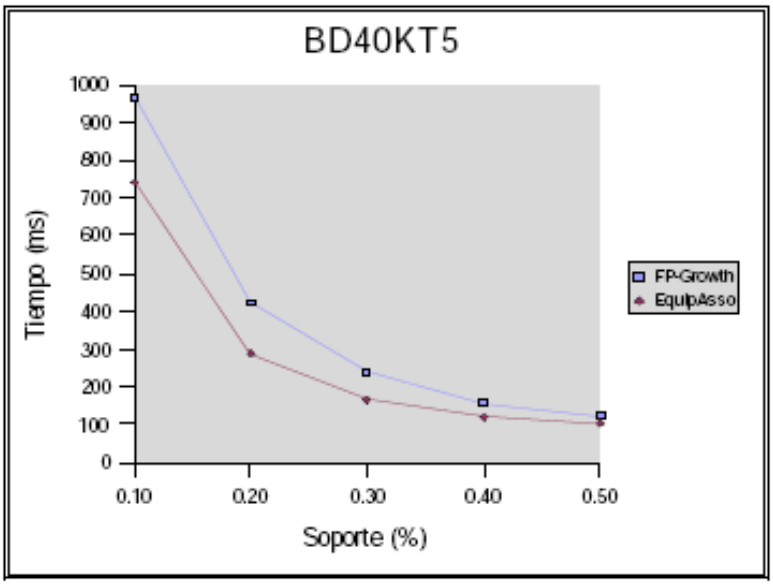

Figura 8.5: Rendimiento BD40KT5

Cuadro 8.7: Tiempos de ejecución tabla BD10KT10

BD10KT10			
Soporte (%)	Tiempo (ms)		
	Apriori	FPGrowth	EquipAsso
0.50	-	257	181
0.75	-	133	93
1.00	-	78	60
1.25	-	61	50
1.50	-	46	42

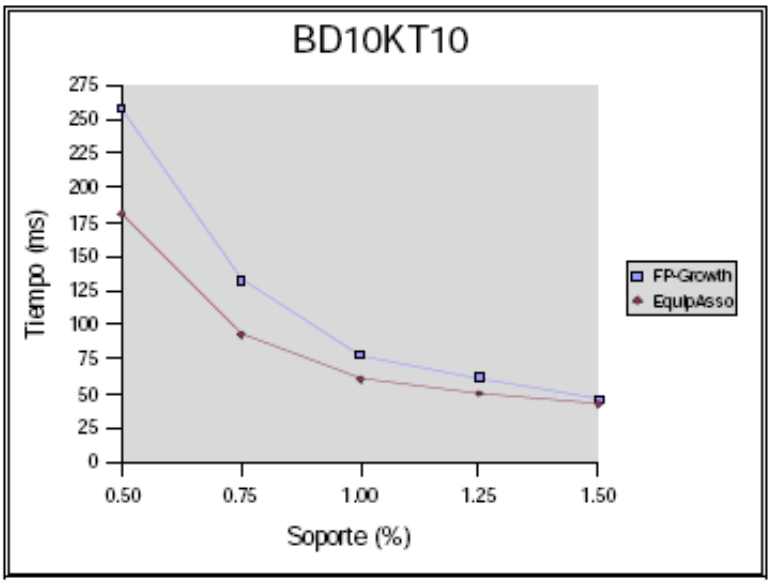

Figura 8.6: Rendimiento BD10KT10

Bibliografía

[1] Université Lumière Lyon 2. Eric equipe de recherche en ingénierie des conaissances. http://chirouble.univ-lyon2.fr, 2006.

[2] NASA Ñational Aeronautics and Space Administration. Tropical cyclone windspeed indicator. http://pm-esip.nsstc.nasa.gov/cyclone/.

[3] R. Agrawal, T. Imielinski, and A. Swami. *Mining Association Rules between Sets of Items in Large Databases*. ACM SIGMOD, 1993.

[4] R. Agrawal, M. Mehta, J. Shafer, R. Srikant, A. Arning, and T. Bollinger. The quest data mining system. In *2Âº Conference KDD y Data Mining*, Portland, Oregon, 1996.

[5] R. Agrawal and K. Shim. Developing tightly-coupled data mining applications on a relational database system. In *The Second International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, 1996.

[6] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *VLDB Conference*, Santiago, Chile, 1994.

[7] R. Brachman and T. Anand. *The Process of Knowledge Discovery in Databases: A First Sketch, Workshop on Knowledge Discovery in Databases*. 1994.

[8] S. Chaudhuri. Data mining and database systems: Where is the intersection? In *Bulletin of the Technical Committee on Data Engineering*, volume 21, Marzo 1998.

[9] M. Chen, J. Han, and P. Yu. Data mining: An overview from database perspective. In *IEEE Transactions on Knowledgeand Data Engineering*, 1996.

[10] Quadrillion Corp. Q-yield. http://www.quadrillion.com/qyield.shtm, 2001.

[11] IBM Corporation. Intelligent miner. http://www-4.ibm.com/software/data/iminer, 2001.

[12] J Demsar and B Zupan. Orange: From experimental machine learning to interactive data mining. Technical report, Faculty of Computer and Information Science, University of Liubliana, Slovenia, http://www.ailab.si/orange/wp/orange.pdf, 2004.

[13] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview, in advances in knowledge discovery and data mining. In *AAAI Pres / The MIT Press*, 1996.

[14] M. Goebel and L. Gruenwald. A survey of data mining and knowledge discovery software tools. In *SIGKDD Explorations*, volume 1 of *1*, June 1999.

[15] Waikato ML Group. Attribute-relation file format (arff). http://www.cs.waikato.ac.nz/ ml/weka/arff.html.

[16] Waikato ML Group. Collections of datasets. *http* : *//www.cs.waikato.ac.nz/ml/weka/index*$_d$*atasets.html*.

[17] Waikato ML Group. The waikato environment for knowledge analisys. http://www.cs.waikato.ac.nz/ml/weka.

[18] J. Han, J. Chiang, S. Chee, J. Chen, Q. Chen, S. Cheng, W. Gong, M. Kamber, K. Koperski, G. Liu, Y. Lu, N. Stefanovic, L. Winstone, B. Xia, O. Zaiane, S. Zhang, and H. Zhu. Dbminer: A system for data mining in relational databases and data warehouses. In *CASCON: Meeting of Minds*.

[19] J. Han, Y. Fu, and S. Tang. Advances of the dblearn system for knowledge discovery in large databases. In *International Joint Conference on Artificial Intelligence IJCAI*, Montreal, Canada, 1995.

[20] J. Han, Y. Fu, W. Wang, J. Chiang, K. Koperski, D. Li, Y. Lu, A. Rajan, N. Stefanovic, B. Xia, and O. Zaiane. Dbminer: A system for mining knowledge in large relational databases. In *The second International Conference on Knowledge Discovery & Data Mining*.

[21] J. Han, Y. Fu, W. Wang, J. Chiang, O. Zaiane, and K. Koperski. *DBMiner: Interactive Mining of Multiple-Level Knowledge in Relational Databases*. ACM SIGMOD, Montreal, Canada, 1996.

[22] J. Han and M. Kamber. *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers, 2001.

[23] J. Han and J. Pei. Mining frequent patterns by pattern-growth: Methodology and implications. In *SIGKDD Explorations*, volume 2:14-20, 2000.

[24] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *ACM SIGMOD*, Dallas, TX, 2000.

[25] Java Hispano. El abc de jdbc. http://javahispano.org/tutorials.type.action?type=j2se, 2004.

[26] T. Imielnski and H. Mannila. A database perspective on knowledge discovery. In *Communications of the ACM*.

[27] RuleQuest Research Inc. C5.0. http://www.rulequest.com, 2001.

[28] E-Business Technology Institute. E-business technology institute, the university of hong kong. http://www.eti.hku.hk, 2005.

[29] Quinlan J.R. *C4.5: Programs for Machine Leraning*. Morgan Kaufmann Publishers, 1993.

[30] Kdnuggets. http://www.kdnuggets.com/software, 2001.

[31] C. Matheus, P. Chang, and G. Piatetsky-Shapiro. Systems for knowledge discovery in databases. In *IEEE Transactions on Knowledge and Data Engineering*, volume 5, 1993.

[32] I Mierswa, M Wurst, R Klinkenberg, M Scholz, and T Euler. Yale: Rapid prototyping for complex data mining tasks. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), 2006.

[33] Faculty of Computer and Slovenia Information Science, University of Liubliana. Orange, fruitful and fun. http://www.ailab.si/orange, 2006.

[34] Faculty of Computer and Slovenia Information Science, University of Liubliana. Orange's interface to mysql. http://www.ailab.si/orange/doc/modules/orngMySQL.htm, 2006.

[35] Goverment of Hong Kong. Innovation and technology fund. http://www.itf.gov.hk, 2006.

[36] Artificial Inteligence Unit of the University of Dortmund. Artificial inteligence unit of the university of dortmund. http://www-ai.cs.uni-dortmund.de, 2006.

[37] G. Piatetsky-Shapiro, R. Brachman, and T. Khabaza. *An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications*. 1996.

[38] J.R. Quinlan. *Induction of decision trees. Machine Learning.* 1986.

[39] R Rakotomalala. Tanagra. In *TANAGRA: a free software for research and academic purposes*, volume 2, pages 697–702. EGC'2005, 2005.

[40] R Rakotomalala. Tanagra project. http://chirouble.univ-lyon2.fr/ ricco/tanagra/en/tanagra.html, 2006.

[41] Isoft S.A. Alice. http://www.alice-soft.com/html/prod_alice.htm, 2001.

[42] S. Sarawagi, S. Thomas, and R. Agrawal. Integrating association rule mining with relational database systems: Alternatives and implications. In *ACM SIGMOD*, 1998.

[43] SPSS. Clementine. http://www.spss.com/clementine, 2001.

[44] Information Technology The University of Alabama in Huntsville and Systems Center. Adam 4.0.2 components. http://datamining.itsc.uah.edu/adam/documentation.html.

[45] Information Technology The University of Alabama in Huntsville and Systems Center. Algorithm development and mining system. http://datamining.itsc.uah.edu/adam/index.html.

[46] R. Timarán. Arquitecturas de integración del proceso de descubrimiento de conocimiento con sistemas de gestión de bases de datos: un estado del arte, en revista ingeniería y competitividad. *Revista de Ingeniería y Competitividad, Universidad del Valle*, 3(2), Diciembre 2001.

[47] R. Timarán. Descubrimiento de conocimiento en bases de datos: Una visión general. In *Primer Congreso Nacional de Investigación y Tecnología en Ingeniería de Sistemas*, Universidad del Quindío, Armenia, Octubre 2002.

[48] R. Timarán. *Nuevas Primitivas SQL para el Descubrimiento de Conocimiento en Arquitecturas Fuertemente Acopladas con un Sistema Gestor de Bases de Datos.* PhD thesis, Universidad del Valle, 2005.

[49] R. Timarán and M. Millán. Equipasso: an algorithm based on new relational algebraic operators for association rules discovery. In *Fourth IASTED International Conference on Computational Intelligence*, Calgary, Alberta, Canada, July 2005.

[50] R. Timarán and M. Millán. Equipasso: un algoritmo para el descubrimiento de reglas de asociación basado en operadores algebraicos. In *4Âª Conferencia Iberoamericana en Sistemas, Cibernética e Informática CICI 2005*, Orlando, Florida, EE.UU., Julio 2005.

[51] R. Timarán, M. Millán, and F. Machuca. New algebraic operators and sql primitives for mining association rules. In *IASTED International Conference Neural Networks and Computational Intelligence*, Cancun, Mexico, 2003.

[52] I. Waitten and F. Eibe. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. 2001.

[53] YALE. Yale - yet another learning environment. http://rapid-i.com, 2006.