

Tópicos Avanzados en Bases de Datos

Profesor: Carlos Hurtado Larrain
(churtado@dcc.uchile.cl)

1 Objetivo

El curso se centra en el área de minería de datos y su objetivo es introducir al alumno en los conceptos y técnicas fundamentales dentro de esta área. Este semestre, el trabajo del curso se centrará fuertemente en un proyecto, en el cuál los alumnos deberán desarrollar un sistema de minería de datos que opere sobre una base de datos que proveerá el profesor.

2 Requisitos

Bases de Datos (42a).

3 Programa

1. **Introducción:** nociones generales, el proceso de minería de datos, extracción, limpieza y transformación de datos.

Referencias¹:

- Themistoklis Palpanas. Knowledge Discovery in Data Warehouses. Sigmod Record 29(3), September 2000.
- A. Maydanchik. Challenges of Efficient Data Cleansing (DM Review - Data Quality resource portal).
- Vijayshankar Raman and Joe Hellerstein, Potters Wheel. An Interactive Framework for Data Cleaning and Transformation (Working Draft 2001)

¹Todos los artículos mencionados en las referencias están disponibles en la Web.

- Helena Galhardas, Daniela Florescu, Dennis Shasha, Eric Simon and Cristian Saita Declarative Data Cleaning: Model, Language, and Algorithms (INRIA Technical Report RR-4149, 2001)
2. **Búsqueda de asociaciones:** algoritmos básicos (Apriori, Apriori-TID, PCY), búsqueda de asociaciones a múltiples niveles, asociaciones vs. correlaciones.

Referencias:

- R. Agrawal, T. Imielinski, A. Swami. Mining Associations between Sets of Items in Massive Databases. Proc. of the ACM-SIGMOD 1993 Int'l Conference on Management of Data, Washington D.C., May 1993, 207-216.
 - R. Agrawal, R. Srikant. Fast Algorithms for Mining Association Rules. Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, Sept. 1994.
 - R. Srikant, R. Agrawal. Mining Generalized Association Rules. Proc. of the 21st Int'l Conference on Very Large Databases, Zurich, Switzerland, Sep. 1995.
 - Sergey Brin, Rajeev Motwani, Craig Silverstein: Beyond Market Baskets: Generalizing Association Rules to Correlations. SIGMOD Conference 1997.
3. **Clasificación:** inducción de árboles de decisión, clasificación Bayesiana, estimación de precisión en clasificación.

Referencias:

- J.C. Shafer, R. Agrawal, M. Mehta. SPRINT: A Scalable Parallel Classifier for Data Mining. Proc. of the 22th Int'l Conference on Very Large Databases, Mumbai (Bombay), India, Sept. 1996.
- M. Mehta, R. Agrawal and J. Rissanen. SLIQ: A Fast Scalable Classifier for Data Mining. Proc. of the Fifth Int'l Conference on Extending Database Technology, Avignon, France, March 1996.
- S. Murthy. Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. Journal of Data Mining and Knowledge Discovery, vol. 2, num. 4, 1998.
- Y. Shih. Families of splitting criteria for classification trees. Statistics and Computing, vol. 9, pp. 309-315, 1999.

- T. Lim, W. Loh, Y. Shih. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning 2000*, vol. 40.
- Minos Garofalakis, Dongjoon Hyun, Rajeev Rastogi, and Kyuseok Shim. Efficient Algorithms for Constructing Decision Trees with Constraints. *Proceedings of ACM SIGKDD'2000*, Boston, Massachusetts, August 2000.
- D. Heckerman. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, March, 1995 (revised November, 1996).

4. **Búsqueda de agrupaciones (Clustering):** tipos de datos en búsqueda de agrupaciones, algoritmos, problemas de segmentación.

Referencias:

- S. Guha, R. Rastogi and K. Shim. CURE: An efficient algorithm for clustering large databases . In *Proceedings of ACM-SIGMOD 1998 International Conference on Management of Data*, Seattle, 1998.
- Rakesh Agrawal, Johannes Gehrke: Dimitrios Gunopulos, Prabhakar Raghavan. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. *Proc. of the ACM SIGMOD Int'l Conference on Management of Data*, Seattle, Washington, June 1998.
- P.S. Bradley, Usama Fayyad, Cory Reina. *Scaling Clustering Algorithms to Large Databases*. *Knowledge Discovery and Data Mining*, 1998.
- J. Kleinberg, C. Papadimitriou, P. Raghavan. Segmentation problems: A micro-economic view of data mining. *Proc. 30th ACM Symposium on Theory of Computing*, 1998.
- Anthony K.H. Kung, Raymond T. Ng, Laks V.S. Lakshmanan, and Jiawei Han. Constraint-based Clustering in Large Databases , *Int. Conf. on Database Theory (ICDT'01)*, London, England, January 2001.

5. **Sistemas de Recomendación y Filtro colaborativo.** Sistemas basados en memoria vs. sistemas basados en modelo. Experiencias y comparaciones experimentales.

Referencias:

- John S. Breese, David Heckerman, Carl Kadie. Empirical Analysis of Predictive Algorithms for Collaborative Filtering (1998).
- R.D. Lawrence, G.S. Almasi, V. Kotlyar, M.S. Viveros, and S.S. Duri. Personalization of Supermarket Product Recommendations. IBM Research Report.
- Analysis of Recommendation Algorithms for E-Commerce (2000), by Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl (GroupLens Research Group / Army HPC Research Center, Department of Computer Science and Engineering University of Minnesota)
- Application of Dimensionality Reduction in Recommender System (2000), by Badrul M. Sarwar, George Karypis, Joseph A. Konstan, John T. Riedl (GroupLens Research Group Army HPC Research Center Department of Computer Science and Engineering University of Minnesota).

6. **Minería de datos en la Web:** minería de uso y estructura de la Web.

Referencias:

- S. Brin, R. Motwani, L. Page, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Stanford Digital Libraries Working Paper, 1998.
- S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. WWW7/Computer Networks (1-7), 1998.
- S. Brin. Extracting Patterns and Relations from the World Wide Web. WebDB Workshop at EDBT'98.
- J. Kleinberg. Authoritative sources in a hyperlinked environment. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998. Extended version in Journal of the ACM 46(1999).
- S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins. Mining the link structure of the World Wide Web. IEEE Computer, August 1999.
- R. Kosala and H. Blockeel. Web Mining Research: A Survey. In SIGKDD Explorations, 2(1):1-15, July 2000.
<http://www.acm.org/sigs/sigkdd/explorations/issue2-1/kosala.pdf>.

- R. Cooley, M. Deshpande, J. Srivastava, P. Tan. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, Vol. 1, Issue 2, 2000.
- M. Deshpande, B. Mobasher, J. Srivastava. Automatic Personalization Based on Web Usage Mining. Communications of the Association of Computing Machinery (CACM), August 2000, pp. 142-151.

4 Método

Curso de 10 UD; dos clases a la semana de una hora y media cada una. El alumno deberá rendir un control escrito (25% de la calificación final) y deberá presentar un artículo técnico listado en las referencias del curso (25% de la calificación final). El resto del trabajo del curso se concentra en un proyecto (50% de la calificación final) en el cual el alumno deberá aplicar una técnica de minería de datos sobre un conjunto de datos sintético o real. El profesor definirá el el conjunto de datos para los proyectos. El proyecto tiene tres partes:

1. Presentación propuesta del proyecto, descripción de herramienta (software) y de conjunto de datos a utilizar.
2. Presentación estado de avance del proyecto.
3. Reporte con los resultados del proyecto.
4. Presentacion final

Para los alumnos que toman el curso como trabajo dirigido la calificación final consiste en la presentación de un artículo técnico (25% de la nota final) y el proyecto (75% de la nota final).

5 Referencias Generales

- Libros:
 - J. Han and M. Kamber. Data Mining Concepts and Techniques. Morgan Kaufman Publishers, 2001.

- D. Hand, H. Mannila, and P. Smyth. Principles of Data Mining. MIT Press, August 2001.
- T. Mitchell. Machine Learning. McGraw-Hill, 1997.
- Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.
- Artículos en la Web:
 - Knuggets, Data Mining, Web Mining & Knowledge Discovery. <http://www.kdnuggets.com/>.
 - NECI Scientific Literature Digital Library. <http://www.researchindex.com/cs>.
- Herramientas (software) en la Web: <http://www.kdnuggets.com/software/>