APLICACIÓN DE MINERÍA DE DATOS PARA FACILITAR EL TRATAMIENTO DE LAS NORMAS DE PRODUCCIÓN DE ATRIBUTOS SEMÁNTICOS EN IDIOMA ESPAÑOL

Lucía Rosario Malbernat; Cecilia Ana Ruz; Silvia Adriana Cobialca
Grupo DM-AS, Departamento de Sistemas, Universidad CAECE
[1]Sub sede Mar del Plata, [2]Sede CABA
lmalbernat@ucaecemdp.edu.ar; cruz@caece.edu.ar; scobialca@caece.edu.ar

RESUMEN

Las Normas de Producción de Atributos consisten en un registro de aquellos aspectos compartidos de la memoria semántica referidos a la forma en que se definen los conceptos. La memoria semántica contiene información compartida por una comunidad de hablantes que les permite comunicarse.
La recolección de normas de producción de atributos semánticos ha proporcionado el insumo para la comprensión de numerosos fenómenos en el campo del estudio de la organización conceptual y la literatura científica señala la necesidad de contar con normas de atributos empíricamente derivados apropiadas para cada comunidad lingüística.
La recolección de las normas genera un volumen de datos muy rico en relaciones y correlaciones que, tratado con técnicas propias de la minería de datos pueden aportar reglas, patrones, tendencias y predicciones útiles para comprender las normas de producción de atributos y de interés para la Psicología Cognitiva y la Neuropsicología.
Se comparte en este trabajo la memoria técnica de un Proyecto de investigación recientemente aprobado en el marco de un Convenio de Cooperación en Investigación entre el Instituto de Psicología Básica, Aplicada y Tecnología. (IPSIBAT) de la Universidad Nacional de Mar del Plata que desde hace años define Normas de Producción de Atributos y la Universidad CAECE.

***Palabras clave*:** *Data Mining*, *Big Data*, tendencias; patrones; predicciones; segmentaciones; clasificaciones; atributos semánticos; normas de producción; lenguaje natural

CONTEXTO

El Proyecto que se reporta, "Aplicación de Minería de Datos para facilitar el tratamiento de las normas de producción de Atributos Semánticos en idioma español", ha sido aprobado por R.R. 388/17 para el período 2018-2019 con fecha de inicio de actividades 1/04/2018 y tiene carácter inter sedes ya que participan investigadores de la Sede Central y la Subsede Mar del Plata.
Está radicado en el Departamento de Sistemas de la Universidad CACE, donde se dictan entre otras, las Carreras Licenciatura e Ingeniería en Sistemas y Licenciatura en Gestión de Sistemas y Negocios y ha presentado recientemente para su aprobación una Maestría en Ciencias de datos e Innovación empresarial. Se va a desarrollar en el marco de actividades conjuntas en Investigación que se vienen realizando desde hace años, de acuerdo con convenios específicos, con el Centro de Investigación en Procesos Básicos, Metodología y Educación (CIMEPB) del Instituto de Psicología Básica, Aplicada y Tecnología (IPSIBAT) de la Universidad Nacional de Mar del Plata. Se propone analizar, procesar y modelar los datos generados por el grupo de investigación en Psicología Cognitiva y Educacional, especializado en la producción de normas de producción de atributos semánticos para el idioma español.
Entre las actividades conjuntas realizadas entre ambas instituciones se cuenta con la experiencia específica de haber aportado precedentemente *know how* y procesamiento

de datos, a través de técnicas específicas de minería de datos, mediante el proyecto, también radicado en el Departamento de Sistemas de la Universidad CAECE, "Aplicación de técnicas de *Data Mining* en gestión de docentes de educación superior" (DM-ES), aprobado por Resolución 549/13. El proyecto generador de los datos para dicho proyecto, radicado en el CIMEPB, era "Competencias para la innovación docente en enseñanza superior: preparación y actitud para el uso de las TIC" (Código 15/H2015; código de subsidio PSI221/14).

1. INTRODUCCIÓN

Las normas de producción de atributos semánticos consisten en colecciones empíricas de las características que las personas utilizan para describir cada concepto. Los datos se obtienen mediante una tarea de generación de propiedades que se logra pidiendo a los participantes, en un experimento controlado, que enumeren las características que mejor describen un cierto conjunto de conceptos. Esta tarea y las normas resultantes son relevantes en diversas áreas de la psicología, y se han utilizado durante décadas para resolver problemas teóricos y prácticos, pero solo recientemente algunas normas se hicieron públicamente disponibles [RVH04].

Establecer las normas de producción de atributos para un concepto permite obtener información cuantiosa acerca de las características y las relaciones de los conceptos y sus atributos para una población particular. Dada la importancia de los atributos semánticos para las teorías sobre memoria semántica, los investigadores han reconocido el valor de colectar normas de producción de atributos para construir modelos, testear hipótesis, disponer de estímulos experimentales y generar tareas de evaluación en el ámbito clínico [GRA04].

El volumen de datos generado en la recolección de normas de producción de atributos semánticos es cuantioso ya que implica interrogar a centenares de individuos sobre centenares de concepto. Dan cuenta de dicho volumen, por ejemplo, la muestra tomada en el marco del Proyecto "Normas de Producción de Atributos Semánticos en Español Rioplatense en Adultos Mayores (Parte I) - 15/H247-" aprobado en 2017, conformada, tal como se describe en el artículo "*Spanish semantic feature production norms for 400 concrete concepts*" [VVC17] por 810 participantes de entre 20 y 40 años (n=324.000). De igual forma, la muestra con la que se trabajó durante la ejecución del Proyecto "Normas de producción de atributos semánticos en castellano rioplatense para un conjunto extenso de objetos vivos y no vivos." (15/H178 , aprobado en 2011) referidos en el artículo "Distribución de los atributos semánticos en función del tipo de categoría y descripción del campo semántico" [VCG11] que reporta los resultados parciales de un sistema de normas de generación de atributos recogidos de 800 participantes sobre 400 conceptos referidos tanto a objetos vivos como a no vivos (n=320.000).

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

El Proyecto que se reporta en este trabajo tiene por objetivo **investigar la estructura** de los datos histórico recolectados por el grupo de investigación Psicología Cognitiva y Educacional para definir las normas de producción de atributos semánticos en idioma español, tomando la descripción de las características de las variables que manejan, dando cuenta, si corresponde, de la existencia de **errores o de datos atípicos** en la recolección de datos y de analizar la existencia de **posibles relaciones entre las variables** para finalmente aportar, -mediante la aplicación de técnicas funcionales y estructurales multivariantes y otras técnicas supervisadas y no supervisadas propias de la extracción de conocimiento y del procesamiento del lenguaje natural-, **predicciones, asociaciones, clasificaciones** o **segmentaciones de los datos** con el fin de resumir y visualizar la información de manera que se facilite la **identificación de tendencias o patrones** que los subyacen.

3. RESULTADOS ESPERADOS

Se ha propuesto recolectar todas las muestras que el Grupo de Investigación Psicología Cognitiva y Educacional viene tomando desde el inicio de sus proyectos vinculados con las normas de producción de atributos semánticos y en base a ellos **diseñar un modelo de datos** apropiado para el tratamiento de los datos mediante recursos de bases de datos y realizar a los datos recolectados todas las **transformaciones y recodificaciones** necesarias a modo de preparación inicial para su posterior análisis.
Se espera, así, poder **describir las características de las variables individuales** mediante distribuciones de frecuencia, medidas de tendencia central y medidas de variabilidad mediante estadística tradicional, analizar la posible **existencia de errores, datos ausentes o atípicos**, analizar la estructura interna de los datos mediante un análisis multivariado que permita detectar **tendencias y patrones**, aplicar técnicas de *Data Mining* para desarrollar reglas combinando las características, aplicando árboles de decisión, clasificaciones y segmentaciones y finalmente **armar un modelo predictivo** para predecir los conceptos.

4. FORMACIÓN DE RECURSOS HUMANOS

El grupo de investigación en Minería de Datos está integrado por profesores del Departamento de Sistemas que toman como insumo académico lo producido en el marco de los Proyectos de Investigación que desarrollan en el tema.

Anualmente se convoca a los estudiantes avanzados del Departamento para integrarse a los Proyectos de investigación aprobados. Recientemente, en el marco de las actividades del Grupo de Investigación en Minería de Datos, dos estudiantes han aprobado su Trabajo Final de Ingeniería (título del trabajo: "Evaluación de Validez de Esquemas de *Clustering* por Medio de la Aplicación de Indices de Bondad"). Se espera que este año los estudiantes consideren como opción este nuevo Proyecto que el Grupo les pone a disposición.

5. BIBLIOGRAFIA

[RVH04] Romero, C. Ventura, S & Hervás, C. Descubrimiento de Reglas de Predicción en Sistemas de e-learning utilizando programación genética. En: R. Giráldez, J. Riquelme & J. Aguilar-Ruiz. (Eds.) Tendencias de la Minería de Datos en España. Red Española de Minería de Datos 1 España, 2004

[GRA04] Giráldez, R. Riquelme, J. & Aguilar-Ruiz, J. (Eds.) Tendencias de la Minería de Datos en España. Red Española de Minería de Datos. TIC2002-11124-E (ISBN 84-688-8442-1), 2004

[RVP10] Romero C., Ventura, S. Pechenizkiy, M. & Baker R. *Handbook of Educational Data Mining*. Chapman and Hall/CRC Press, Taylor & Francis Group. ISBN: 9781439804575, 2010

[KR02] Kimball, R. Ross, M. *The Data Warehouse Toolkit. The Complete Guide to Dimensional Modeling* (2a ed.) USA: Wiley Publishing, Inc, pp. 243-254, 2002

[ARG09] Argibay, J.C. Muestra en investigación cuantitativa. Revista Subjetividad y Procesos Cognitivos, ISSN 1666-244X, Nº. 13 p. 18, 2009

[GRQ09] Garcia Aretio, L. Ruiz Corbella, M. Quintanal Díaz, J. García Blanco, M. & García Perez, M. Concepción y Tendencias de la Educación a Distancia en América Latina. Colección Documentos de Trabajo. Centro de Altos Estudios Universitarios de la OEI, p. 44, 2009

[VEG11] Vega Pons, S. Combinación de resultados de Clasificadores no supervisados. Tesis de doctorado. Rep. Téc. Reconocimiento de Patrones. Serie Azul. Cuba: Centro de Aplicaciones de Tecnologías de Avanzada, pp. 13, 47. 2011

[YBP09] Yolis, E. Britos, P. Perichisky, G. & García-Martínez, R. Algoritmos Genéticos Aplicados a la Categorización Automática de Documentos. Revista Electrónica de sistemas de Información. ISSN 1677-3071 Doi: 10.5329/RESI, 2, 2009

[BF98] Bradley P.S. & Fayyad, U.M. *Refining initial points for k-means clustering. In J. Shavlik, editor, Proceedings of the Fifteenth International Conference on Machine Learning* (ICML '98), San Francisco, CA, 1998. Morgan Kaufmann, pp. 1-2, 91-99, 1998

[NOR13] Northii, M. *Data Mining for the Masses. A Global Text Project Book* ISBN: 0615684378, 2013. Disponible en: http://docs.rapid-i.com/files/DataMiningForTheMasses.pdf

[RAP12] How to Extend RapidMiner 5White Paper Rapid-I GmbH, 2012

[PSP12] *PSPP Users' Guide GNU PSPP* Statistical *Analysis Software Release* 0.8.0-g13bf3f, 2012. Disponible en: http://www.gnu.org/software/pspp/manual/pspp.html

[BFH13] Bouckaert, R.R. Frank, E. Hall, M. Kirkby, R. Reutemann, P. Seewald, A. Scuse, D. WEKA *Manual for Version* 3-7-10, University of Waikato, Hamilton, New ZealandJuly 31, 2013

[MVM17] MacIntyre, M. Vivas, L Vivas, J. *Tipologia de atributos ponderada baseadas em normas de produção atributos semánticos*. Temas em Psicologia 25 (2), 843-854, 2017

[VVC17] Vivas, J. Vivas, L. Comesaña, A. Coni, A.G. Vorano. *Spanish semantic feature production norms for 400 concrete concepts. Behavior research methods* 49 (3), 1095-1106, 2017

[VMR15] Vivas, J. MacIntyre, M. Ricci, L. Vivas, J. *Psycholinguistic variables involved in concept recall from the successive presentation of features.* Estudios de Psicología 36 (3), 592-619, 2015

[CV15] Comesaña, A. Vivas, J. Evolución de la categorización semántica en adultos mayores con diagnóstico de DCL-A y DTA y sin patología neurológica Interdisciplinaria 32 (1), 7-29, 2015

[CSA14] Cervigni, M Sguerzo, M.R. Alfonso, G. Pastore, M. Martino, P. Mazzoni, C. *Bibliometric analysis of empirical studies in Spanish on Working Memory* (1999-2014), 2014

[ZV15] Zapico, M. Vivas, J. La sinonimia desde una perspectiva linguístico-cognitiva. Medición de la distancia semantic. Onomázein, 198-211, 2015

[HLV15] Huapaya, C. R. Lizarralde, F.A.J. Vivas, J. Modelo para visualizar y evaluar el conocimiento conceptual. TE & ET 2015

[FCM14] Favarotto, V Coni, A.G. Magani, F. Vivas, J. *Semantic Memory Organization. In Children And Young Adults*. Procedia-Social and Behavioral Sciences 140, 92-97, 2014

[ZV14] Zapico, M. Vivas, J. *La sinonimia como caso particular de distancia semantic. Encontros Bibli: revista eletrônica de biblioteconomia e ciência* da, 2014

[LVV12] Lamas, V Vivas, J. Vorano, A. Comparación de atributos semánticos entre diferentes lenguas. IV Congreso Internacional de Investigación y Práctica Profesional en Psicología, 2012

[MVV12] Morales, F. Vivas, J. Vorano, A. Intyre, M. Damian, M. Campoy, P. Normas de Producción de Atributos Semánticos: Diferencias de acuerdo a la edad. IV Congreso Internacional de Investigación y Práctica Profesional en Psicología, 2012

[VCG11] Vivas, J. Comesaña, A García Coni, A Vivas, L. Yerro, M. Distribución de los atributos semánticos en función del tipo de categoría y campo semántico. Resultados preliminares para la confección de normas de atributos. M.C. Richaud y V. Lemos (comp.) Psicología y otras ciencias del comportamiento, 2011

[PYF11] Pazgón, E. Yerro Avincetto, M Favarotto, V. Vivas, L. Vivas, J. Categorización de rasgos semánticos: Diferencias de género en una tarea de atributos de conceptos. Perspectivas en Psicología: Revista de Psicología y Ciencias Afines 8 (2), 2011